AD_______________

Award Number: W81XWH-04-1-0064

TITLE: Mechanisms and Chemoprevention of Ovarian Carcinogenesis

PRINCIPAL INVESTIGATOR: Dusica Cvetkovic, Ph.D.

CONTRACTING ORGANIZATION: Fox Chase Cancer Center
Philadelphia, PA 19111

REPORT DATE: February 2009

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE 01-02-2009	**2. REPORT TYPE** Final	**3. DATES COVERED** 1 FEB 2004 - 31 JAN -2009

4. TITLE AND SUBTITLE Mechanisms and Chemoprevention of Ovarian Carcinogenesis	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER W81XWH-04-1-0064
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Dusica Cvetkovic, Ph.D.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
Email: di_cvetkovic@fccc.edu	**5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Fox Chase Cancer Center Philadelphia, PA 19111	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Ovarian cancer is the most fatal gynecological malignancy. The understanding of the early molecular events leading to ovarian cancer is important for the development of strategies for early detection and prevention. We have demonstrated that DMBAinduced mutagenesis in the rat ovary, combined with gonadotropin hormone-mediated enhanced mitogenesis of the ovarian surface epithelium, produces lesions ranging from preneoplastic, early neoplastic to advanced ovarian tumors, resembling human disease. The goal of this project was to use the DMBA-gonadotropin animal model to study molecular mechanisms underlying ovarian oncogenesis and to conduct a preclinical chemoprevention trial. The original specific aims of the study were: 1) Determine the molecular genetic mechanisms underlying ovarian oncogenesis in the rat DMBA/gonadotropin model of ovarian cancer; 2) Determine the efficacy of the COX-1 inhibitor SC-560 to prevent the appearance and/or progression of DMBA-induced ovarian lesions; and 3) Study the in vivo mechanisms of the putative chemopreventive action of COX-1 inhibition. However, due to change of Principal Investigator in the last year of the study, the original research plan has been modified. Since the animal protocol pertaining to this project has been closed and the proposed chemoprevention trial in rats has not been initiated, only aim 1 is being carried out.

15. SUBJECT TERMS
ovarian carcinogenesis, animal models and cDNA microarrays

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	38	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

MECHANISMS AND CHEMOPREVENTION OF OVARIAN CARCINOGENESIS
FINAL PROGRESS REPORT

INTRODUCTION

Ovarian cancer is the most fatal gynecological malignancy because of its asymptomatic development and frequent diagnosis at an advanced stage. The understanding of the early molecular events leading to the disease is important for the development of strategies for its early diagnosis and prevention, which could improve patient survival and quality of life. We have demonstrated that DMBA-induced mutagenesis in the rat ovary, in combination with gonadotropin hormone-mediated enhanced mitogenesis of the ovarian surface epithelium, produces lesions ranging from preneoplastic, early neoplastic to advanced ovarian tumors, which resemble human disease. The goal of this research project was to use the DMBA-gonadotropin animal model to study the molecular mechanisms underlying ovarian oncogenesis and to conduct a preclinical trial for its chemoprevention. The original specific aims of the study were:

1) Determine the molecular genetic mechanisms underlying ovarian oncogenesis in the rat DMBA/gonadotropin model of ovarian cancer

2) Determine the efficacy of the COX-1 inhibitor SC-560 to prevent the appearance and/or progression of DMBA-induced ovarian lesions

3) Study the *in vivo* mechanisms of the putative chemopreventive action of COX-1 inhibition

However, due to change of Principal Investigator (PI) in the last year of the study, the original research plan has been modified. Since the animal protocol pertaining to this project has been closed and the proposed chemoprevention trial in rats has not been initiated, only specific aim 1 is being carried out.

BODY

During the course of the project supported by this DOD-CDMRP grant, the following progress has been achieved along the proposed aims of the study:

1) Determine the molecular genetic mechanisms underlying ovarian oncogenesis in the rat DMBA/gonadotropin model of ovarian cancer. A large number of DMBA-induced ovarian lesions were generated in the rat at different stages of neoplastic development to provide statistical power and significance of the findings from their molecular classification and characterization. Using funds provided by the Fox Chase Cancer Center (FCCC) NCI Ovarian Cancer SPORE Grant, a two-phase carcinogenesis experiment was initiated at the end of 2003, in which 160 female 6-week old virgin female Sprague-Dawley rats were subjected to bilateral survival surgery to the ovaries. Animals were divided into four arms and treated: a) Control groups a1 (20 animals, no hormones) and a2 (20 animals, with hormones): beeswax-impregnated surgical sutures were implanted in the portion of each ovary that is contra-lateral to the fallopian tube; b) DMBA-/+hormone group (total 100 animals), b1 DMBA/beeswax-impregnated surgical sutures were implanted bilaterally in the ovaries of the animals as above and b2. Two months following the surgical procedure, rats in group a2 and b2 were subjected to four cycles of sequential administration of hormones PMSG and hCG. These procedures are described in the

Experimental Design and Methods section of our grant proposal and in our Cancer Research paper [1]. All treated animals were maintained for one year from the survival surgical procedure, or until disease development and animal distress became evident. Rats were sacrificed according to the initiation of treatment, in December 2004 and January 2005, following the Institutional Animal Care and Use Committee (IACUC) approved guidelines.

All of the ovaries were harvested and fixed in 70% ethanol at 4°C for 18 hr, paraffin processed through a 12 hr cycle with a Tissue-Tek VIP 5 (Sakura Finetek, Torrance, CA) vacuum infiltration processor, and then paraffin embedded with a Histo-Center II (Fischer Scientific, Pittsburgh, PA) embedding station. Three 5 µm-sections, approximately 50 µm apart of each other were obtained from the two end-portions of each ovary, stained with H&E and subjected to histopathological evaluation.

Table 1 indicates the incidence of ovarian lesions observed in the four experimental arms, subdivided into 3 subgroups (nonneoplastic, putative preneoplastic and neoplastic lesions). This experiment was performed to verify the potential promoting role of gonadotropin hormones in ovarian cancer development, and to generate sufficient numbers of ovarian lesions for molecular characterization and elucidation of the mechanisms behind their development. Based on the observed statistically significant differences in lesion incidence between arms a1 and a2, and b1 and b2 (Table 2), and our published data [2], we conclude that gonadotropin hormones play a major role in the promotion of ovarian cancer.

Table 1. DMBA ovarian carcinogenesis with gonadotropin co-treatment

Experimental Arm	*per ovary*				*per animal*			
	No Lesions	Non-Neoplastic Lesions	Putative Pre-Neoplastic Lesions	Neoplastic Lesions	No Lesions	Non-Neoplastic Lesions	Putative Pre-Neoplastic Lesions	Neoplastic Lesions
a1 - Surgery only (20 animals) %	37.5	40.0	22.5	0.0	0.0	70.0	30.0	0.0
a2 - Surgery+Hormones (19 animals) %	20.8	21.1	58.1	0.0	0.0	26.1	73.9	0.0
b1 - DMBA (47 animals) %	15.7	20.5	62.8	1.0	6.3	13.0	78.7	2.1
b2 - DMBA+Hormones (45 animals) %	1.1	15.4	75.8	7.7	0.0	8.8	75.8	15.4

Table 2. Statistical significance of differences in lesion incidence induced by gonadotropin co-treatment (* - determined by χ-square and/or Fisher's exact tests)

Comparison*	Site of the lesions	P-value
Surgery vs. Surgery+Hormones	**Ovary**	**0.0061**
	Animal	**0.0064**
DMBA vs. DMBA+Hormones	**Ovary**	**0.0002**
	Animal	**0.0422**

A number of different types of histologic changes were observed in the ovary [1]. Nonneoplastic lesions include chronic inflammation, foreign body granuloma, suture granuloma, scar and prominent corpora lutea; whereas putative preneoplastic lesions represent bursal and ovarian surface epithelial (OSE) papillomatosis, real stratified and pseudostratified hyperplasia, inclusion cysts and deep invaginations. All the preneoplastic lesions can present with or without atypia. Both the preneoplastic and neoplastic ovarian lesions in arms b1 and b2 displayed a more complex, advanced histology, such as thicker stratified epithelium and more pronounced papillary structures or surface invaginations, relative to those in arms a1 and a2. The incidence of cancer in the DMBA/gonadotropin rat model of ovarian oncogenesis was 6%. Namely, 8 neoplastic lesions were observed in 131 animals, 7 in arm b2, and one in arm b1, out of which 6 were invasive (an undifferentiated and a differentiated adenocarcinoma, a Leydig-Sertoli tumor, two granulosa/theca cell tumors, and a papillary serous tumor).

A similar report has recently demonstrated that rats treated with systemic estrogen and local ovarian DMBA administration simultaneously develop preneoplastic and neoplastic lesions in the breast and ovary [3]. The same criteria was used to evaluate progression toward ovarian cancer as in our study, namely putative ovarian preneoplastic changes such as inclusion cysts, epithelial hyperplasia, papilloma and stromal hyperplasia.

Molecular characterization of DMBA/gonadotropin-induced rat ovarian lesions

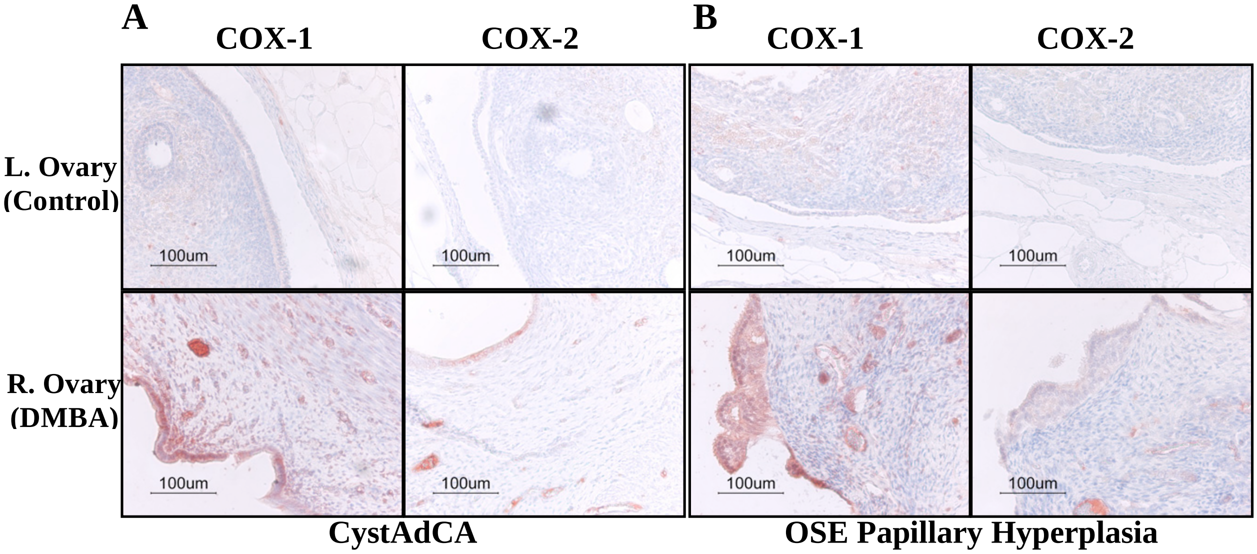

Figure 1. IHC staining for COX-1 (left half-panel A and B) and COX-2 (right half-panel A and B) protein expression in rat ovaries: Left (L. Ovary) untreated control (top panels) and Right (R. Ovary) DMBA-treated (lower panels). **A.** Cystadenocarcinoma; **B.** Surface epithelial papillary hyperplasia. Sections of left and right ovary from the same animal were mounted on the same slide and subjected to IHC at identical conditions. Pictures of each pair of sections per slide were taken at identical brightness/contrast settings.

Tp53 and Ki-Ras point mutations, that are characteristic for human ovarian cancer, are also present in the DMBA/gonadotropin-induced preneoplastic rat ovarian lesions. Additionally, an overexpression of estrogen and progesterone receptors in preneoplastic and early neoplastic lesions and their loss in advanced tumors, suggest a role of these receptors in ovarian cancer development [1].

To determine whether, similar to human disease [4], COX-1 and/or COX-2 expression/activation is linked with ovarian neoplastic development in this animal model, we initiated collaboration with Dr. S. K. Dey at Vanderbilt University Medical Center. Histological slides were prepared from tissue sections obtained from formalin-fixed paraffin-embedded (FFPE) rat ovaries treated with DMBA or DMBA/hormones and containing putative preneoplastic (7 samples) or neoplastic lesions (5 samples). Each slide also contained a tissue section from the corresponding contra-lateral, control ovary. Individual slides, sent to Dr. Dey, were subjected to immunohistochemical (IHC) analysis for COX-1 or COX-2 expression. Elevated expression of both enzymes was observed in the majority of analyzed putative preneoplastic lesions and all neoplastic lesions regardless of progression. Neither protein was detectable in the OSE of normal (control) ovaries. Even though in most cases, the expression level of COX-1 was higher than that of COX-2, the data implied a strong association of both enzymes with ovarian cancer development in this model. Figure 1 shows examples of changes in COX-1/2 expression. These results are interesting, and though they support our original proposal for the pre-clinical testing of a COX-2 specific inhibitor (celecoxib) (see 2. below), they also suggest that a COX-1 specific inhibitor (such as SC-560, Cayman Chemical Co) may be more effective as an agent for chemoprevention of ovarian cancer. The results also warrant further analysis of additional ovarian lesions, both putative preneoplastic and neoplastic, in order to evaluate the prevalence of the observed changes in COX-1/2 expression, and whether they are also present in putative preneoplastic lesions induced by gonadotropin hormone treatment alone.

We have previously performed a global, microarray-based gene expression analysis of human ovarian tumors and normal human ovarian surface epithelia (non-cultured or short-term cultured). Among the genes identified with differential expression between different types of tumors and normal OSE, the most interesting was the NF-κB regulator gene A20. While this gene was found expressed at moderate to high levels in the normal OSE, its expression was undetectable in all tested tumors, irrespective of their histological subtype or neoplastic stage (Fig. 2). This result suggests that A20 plays a confounding role in the development of ovarian carcinomas and could potentially play such a role in the DMBA/gonadotropin model. A20 is an enzyme with dual ubiquitination and de-ubiquitination activities and plays an important role as a switch between activation and inactivation of the NF-κB survival transcription factor [5, 6]. While A20 facilitates the coupling of cytokine and other receptor signals to the IKK signalosome complex through RIP and other MAP3Ks, it is also essential for termination of the same signals and inhibition of a persistent NF-κB activation. The persistent, elevated activation of NF-κB has been associated with the malignant progression and development of resistance to cytotoxic treatment of many types of tumors. Therefore, loss of A20 in ovarian cancer may be one of the underlying mechanisms and a very important target for the design of new strategies for prevention and treatment of the disease. In support of this observation, the preliminary results

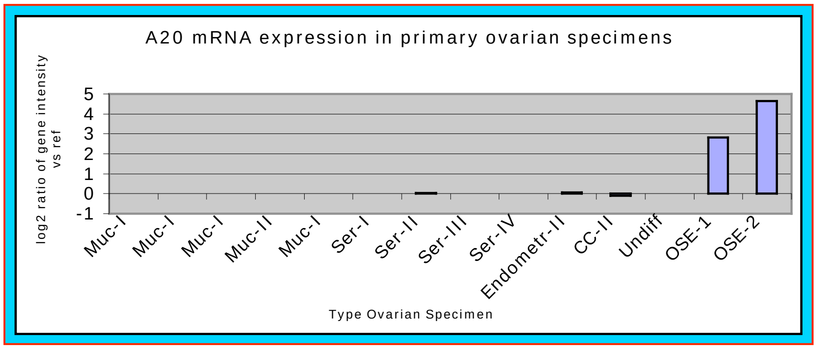

Figure 2. Microarray-determined A20 mRNA expression in primary human ovarian cancer specimens of different histological subtype and malignant stage, and in normal human OSE (OSE-1: average of 4 short-term cultures; OSE-2: average of 2 non-cultured samples). Data was confirmed by real-time qRT-PCR analysis (data not shown)

obtained from a phase I trial of the proteasome inhibitor bortezomib in combination with platinum agents (carboplatin) for overcoming the development of chemoresistance of ovarian cancer patients are encouraging [7]. Based on our results from the analysis of human normal OSE, we suggest that A20 would also be expressed at moderate levels in the normal rat OSE. Though the examination of expression status of A20 in the normal rat OSE and in DMBA/hormone-induced lesions at different stages of neoplasia by real-time qRT-PCR was originally planned, the analysis has not been initiated.

Genomic analysis of DMBA/gonadotropin-induced rat ovarian lesions

With the guidance of our collaborator, pathologist Dr. A. Klein-Szanto, we have achieved a complete histopathological examination of 262 ovaries harvested from 131 animals included in the four arms of the carcinogenesis experiment described above. This allowed the identification of ovaries that contain different types of lesions and the selection of lesions for the purpose of this study according to their classification. In a streamline fashion, ovaries selected for a certain type of lesion were then subjected to further processing in preparation for genomic analysis. In order to better preserve the quality of RNA, ethanol-fixed paraffin-embedded (EFPE) ovarian tissue blocks were kept at 4°C at all times. Depending on the size of lesion and its epithelial cell component, 4-6 5μm-sections were generated from the portion of the organ adjacent to the corresponding H&E sections and either stored at –80°C until they were subjected to laser-capture microdissection (LCM) or processed immediately. Prior to proceeding with laborious microdissections, the quality of isolated RNA was checked on tissue scrapes, using the Agilent 2100 Bioanalyzer and samples with unadequate quality were excluded from the analysis. Ovarian tissue sections were stained with HistoGene LCM Staining Kit (Arcturus /Molecular Devices, Sunnyvale, CA), and 2,000-5,000 cells from DMBA/gonadotropin-induced ovarian lesions were collected on CapSure LCM Caps using either PixCell II or AutoPix LCM Systems (Arcturus). It is estimated that 10 pg of RNA is obtained from a single cell, therefore 5,000 of LCM-captured cells contain approximately 50 ng of RNA.

It has been reported that a considerable variation in the microarray data is incorporated when different sets of arrays are used to compare specimens in a single experiment. To avoid this, and since the preparation of tissue specimens, purification and amplification of RNA and quality testing are the rate-limiting procedures, we have processed all lesion samples to the point where all hybridizations are carried out serially within a short period of time and with the same lot of microarrays.

We would like to emphasize that in February of 2007 the PI status on the project has changed. Dr. Patriotis had left FCCC, and Dr. Cvetkovic, who had no prior involvement in this project, took over to finish up the study. LCM-derived tissue samples generated along the lines of this DOD-funded research were transferred to the new laboratory. However, these samples were fixed by an alternative method, using ethanol, and then paraffin embedded, while the golden standard for molecular analyses are snap-frozen tissue specimens [8, 9]. The rationale behind ethanol fixation was to preserve tissue architecture and cellular morphology of the rat ovary, while allowing for the recovery of good quality RNA from microdissected cells. Despite the loss in morphologic quality in frozen sections, especially in non-cover-slipped slides for LCM, RNA quality is generally much better than RNA obtained from ethanol- or formalin-fixed tissues [10]. Moreover, the Arcturus LCM systems that were initially used to procure biological samples for this study have in the meantime undergone substantial technical improvements. The newer generations of platforms, the upgraded manual PixCell II, and the automated Veritas and Arcturus XT Microdissection Systems, have features that allow for superior visualization of cellular morphology, irrespective of the tissue fixation method, compared to previous generation PixCell II and AutoPix systems.

Though others have successfully recovered RNA from EFPE human and animal tissues sufficient for downstream molecular profiling studies [11, 12], we wanted to check the quality and amplifiability of RNA from DMBA/hormone-induced rat ovarian lesions on several levels prior to microarray analysis. We have consulted with the application scientists at Arcturus on how to approach this issue. Since Arcturus makes kits designed exclusively for extraction of RNA from frozen (PicoPure RNA Isolation Kit) or FFPE tissues (Paradise Reagent System), we needed to determine which one would be more appropriate for our EFPE samples. In addition to these, two other kits were included in the test, Recover All Total Nucleic Acid Isolation Kit (Ambion/Applied Biosystems, Austin, TX) and Optimum FFPE RNA Isolation Kit (Asuragen). Two randomly selected EFPE rat ovarian tissue samples from our experiment where cut onto four slides, and each one was scraped off and used for RNA extraction with one of the four nucleic acid isolation kits. The quantification and integrity determination of isolated RNA were carried out by micro fluidic electrophoresis on Agilent 2100 Bioanalyzer using the RNA 6000 Pico LabChip Kit (Agilent Technologies, Santa Clara, CA). Additional sample quality assessment was done by quantitative real-time PCR using the protocol developed by Arcturus (Paradise Sample Quality Assessment Kit). This protocol utilizes 3' and 5' primer sets to amplify a portion of the beta-actin gene. The 3'/5' ratio evaluates the abundance of the average beta-actin cDNA from the 3' end compared to the abundance of a 5' sequence using the quantified PCR yields of each amplicon. If most of the cDNA contains both the 3' and 5'sequence target, the ratio of the PCR product for 3'/5' is close to one. As the RNA starts exhibiting some level of degradation, the 3'/5' ratio tends to become greater than one. Depending on the ratio, an estimation of the RNA quality can be made. A suggested cut-off is ≤20. Using four different nucleic acid isolation kits, both sample 1 and sample 2 yielded 3'/5' ratios in the range from 3-11 (Table 3), indicating acceptable quality and amplifiability of RNA from DMBA/hormone-induced rat ovarian lesions. There were no significant differences between the four kits; hence we decided to use the PicoPure RNA Isolation Kit, as originally proposed.

RNA from EFPE rat tissue scrapes exhibited in general a heterogeneous profile on the Bioanalyzer, with either broadened 18s and 28s peaks, or without the peaks (Figure 3). These profiles indicate compromised integrity of RNA, more resembling RNA profiles of FFPE tissues, than those of frozen tissues. However, researchers from our and other institutions have successfully performed microarray analysis on partially degraded RNA [13, 14]. Based on published data, we felt that our LCM-derived, partially degraded RNA with relatively low RNA integrity number (RIN) values, would still be viable in microarray analysis.

Table 3. Comparison of RNA isolation kits for EFPE samples

	Optimum FFPE RNA Isolation Kit (Asuragen)	All Total Nucleic Acid Isolation Kit (Ambion)	PicoPure RNA Isolation Kit (Arcturus)	Paradise RNA Isolation System (Arcturus)
Sample 1 3'/5' ratio	2.8	7.0	11.4	2.9
Sample 2 3'/5' ratio	2.7	3.8	6.2	7.8

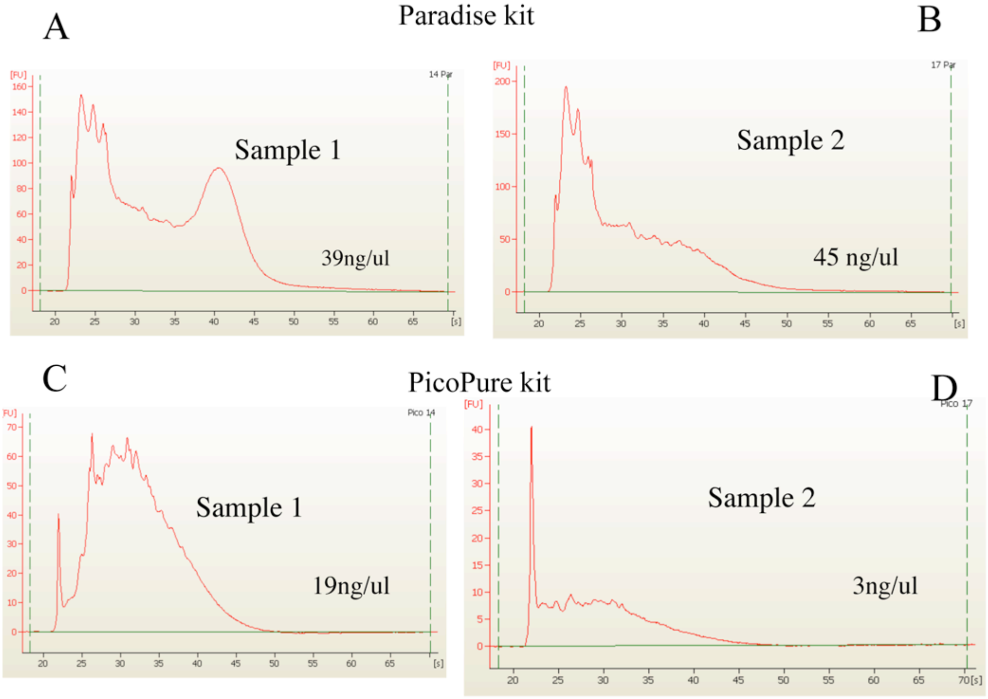

Figure 3. Representative bioanalyzer profiles of RNA isolated from EFPE rat ovaries by Paradise (A, B) and Picopure Kits (C, D) (tissue scrapes)

Amino Allyl MessageAmp II aRNA Amplification Kit (Ambion) was used to amplify and Cy-3-label 24 individual LCM-derived samples, 8 in each of the three above described ovarian lesion categories/groups (nonneoplastic, putative preneoplastic and neoplastic). These samples were from b1 and b2 arms of the experiment. Quantification and integrity assessments of RNA were carried out on the Bioanalyzer. One of the primary limitations of microarray analysis is large amount of labeled input RNA (several µg) required for hybridization [15]. When the starting cell population is limited, such as in LCM-procured samples, a second round of linear amplification is necessary in order to have sufficient quantities of amplified RNA (aRNA) to use for probe synthesis. In our hands, approximately 50 ng of total RNA is amplified in two rounds and 1 µg of Cy3-labeled aRNA is put into hybridization reaction. Universal Rat Reference RNA (Stratagene, La Jolla, CA) is used in the positive control amplification reaction.

Although previous annual reports have indicated the intent to use the Affymetrix GeneChip system for the genomic analysis of rat ovarian lesions, due to change of PI, limited time frame and resources, as well as cost-effectiveness, the decision has been made to utilize the Agilent platform instead. This platform is available at the Fox Chase Cancer Center DNA Microarray Facility. Cy3-labeled samples were hybridized to Agilent 4x44K Whole Rat Genome arrays. Microarray images were processed using Agilent Feature Extraction software, v9.5. RNA sample quality issues and array quality control failures necessitated the removal of several arrays from the analysis, leaving 5 nonneoplastic samples and 6 each from the other two groups, preneoplastic and neoplastic.

Array data was preprocessed and analyzed using Bioconductor's *limma* package [16, 17]. Median signal intensities were background corrected using the *normexp* method, and quantile normalization was performed to make intensity distributions consistent across arrays. Prior to differential expression analysis, a non-specific filter was applied to the probe list: probes were removed if they lacked association with an Entrez gene ID, or if they had expression intensities close to background for a large percentage of the arrays.

Differential expression analysis between all pairs of groups was performed using the *limma* package, which implements the computation of empirical Bayes moderated two-sample t-statistics. P-values from these tests were adjusted for multiple comparisons using the Benjamini-Hochberg method to control the false discovery rate (FDR) [18]. A probe was declared significant if it had a FDR less than 5%. With this significance criterion, there were no differentially expressed probes for the comparisons of preneoplastic vs. nonneoplastic or neoplastic vs. preneoplastic, within or among b1 and b2 arms (Figure 4). Specifically, no changes in gene expression were found in arm b1, between nonneoplastic and preneoplastic samples, and in arm b2, between nonneoplastic and preneoplastic samples; also no changes in arm b1 among preneoplastic and neoplastic, and in b2 among preneoplastic and neoplastic samples. There were 558 probes identified as significantly differentially expressed in the comparison between neoplastic, in either b1 or b2 arms, to its respective nonneoplastic controls. The inherent problem with this study was only one neoplastic/cancer lesion in b1 arm. Therefore, it made sense to analyze the data within the experimental arms.

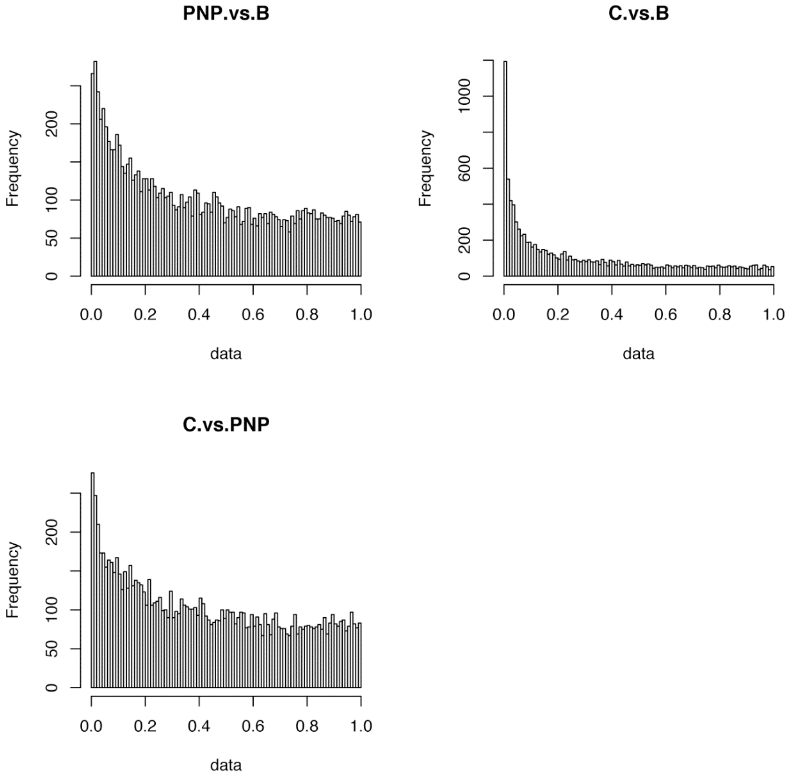

Figure 4. Probability histogram of microarray differences between preneoplastic vs. nonneoplastic (PNP vs B), neoplastic vs. nonneoplastic (C vs B) and neoplastic vs. preneoplastic (C vs PNP) samples

In our microarray analysis of the rat ovarian lesions we expected to identify genes whose changes in expression are associated with increased ovarian lesion severity and malignant progression, from nonneoplastic and preneoplastic to neoplastic. We wanted to determine whether a continuum of OSE cell malignant development exists in this model, similar to the multistep progression model of colorectal tumorigenesis proposed by Fearon and Vogelstein

[19], and to identify genes whose changes in expression and/or functional activity are associated with this process. The apparent OSE cell origin of DMBA-induced tumors [20] make this model not only convenient, but also relevant to disease in women and perhaps valid for testing of new prevention agents. Since we did not observe significant changes in gene expression among nonneoplastic vs. preneoplastic, and among preneoplastic vs. neoplastic lesions, it appears that DMBA/hormone treatment in the rat causes tumor formation without step-wise progression from benign to malignant. Among the total of 558 differentially expressed probes between the neoplastic and nonneoplastic group, we found a number of interesting genes that are associated with human ovarian cancer. We have used a cut-off value of 4-fold for both up and downregulated genes, cancer group versus nonneoplastic control, to shorten the original list of genes (Tables 4 and 5). The most interesting genes in the upregulated group include those encoding for vascular endothelial growth factor A; cholinergic receptor, nicotinic, beta polypeptide 4; tumor suppressors breast cancer 2 and Ras association (RalGDS/AF-6) domain family member 2; two dynamins, dynamin 1-like and dynamin 2; two protein phosphatase associated genes, protein phosphatase 1 (formerly 2C)-like and protein phosphatase 1, regulatory (inhibitor) subunit 9A; cisplatin resistance-associated overexpressed protein; ATP-binding cassette, sub-family B (MDR/TAP), member 1 that is involved in multidrug resistance; a structural protein that predicts prognosis of ovarian cancer in women, procollagen, type IV, alpha 4; and cellular retinoic acid binding protein 1 involved in vitamin A signaling. There is a clinical trial for recurrent ovarian cancer involving anti-VEGF antibody. Some of the interesting genes from the list of downregulated transcripts are insulin growth factor l; collagen, type I, alpha 2; cell adhesion associated cadherin, EGF LAG seven-pass G-type receptor 2 (flaming), and fatty acid binding protein 3, muscle and heart. These genes have been studied in human ovarian cancer via microarray and other types of analyses [21-25]. It is interesting that our microarray analysis did not show differences in the expression among groups of hormone receptors, as suggested by our IHC results.

Table 4. Genes upregulated in neoplastic vs. nonneoplastic rat ovarian lesions (>4-fold), associated with human ovarian cancer

Gene	**Ref Seq**	**Fold change⇑**	**FDR**
Chrnb4	NM_052806	43.45	0.013
Dnm2	NM_013199	30.15	0.015
Col4a4	NM_001008332	20.54	0.021
Tpm3	NM_057208	19.65	0.021
Erbb2	NM_017003	11.00	0.021
Dnm1l	NM_053655	10.75	0.015
Vegfa	NM_001110333	9.99	0.023
Ppm1l	NM_001107681	8.56	0.039
Brca2	NM_031542	7.73	0.020
Hnrnpa1	NM_017248	7.59	0.013
Rassf2	NM_001037096	7.42	0.032
Csnk1a1	NM_053615	6.38	0.020
Hdac5	XM_001081495	6.20	0.017
Rab8a	NM_053998	6.19	0.020
Ppp1r9a	NM_053473	5.47	0.030
Smptb	NM_182818	5.46	0.038

Arrb1	NM_012910	5.34	0.038
Crop	NM_001108291	5.26	0.024
Abcb10	NM_001012166	5.20	0.019
Hras	NM_001098241	5.16	0.045
Car5a	NM_019293	4.68	0.038
Arnt2	NM_012781	4.52	0.032
Npm1	NM_012992	4.34	0.015
Arhgap10	NM_001109501	4.31	0.043
Gadd45b	NM_001008321	4.30	0.013
Crabp1	NM_001105716	4.28	0.022
Plxna2	NM_001105988	4.25	0.043

Table 5. Genes downregulated in neoplastic vs. nonneoplastic rat ovarian lesions (>4-fold), associated with human ovarian cancer

Gene	**Ref Seq**	**Fold change⇓**	**FDR**
Crhr1	NM_030999	4.31	0.032
Igf1	NM_001082477	4.45	0.42
Btbd3	NM_001107782	4.60	0.021
Col1a2	NM_053356	4.87	0.032
Ercc6	NM_001107296	5.25	0.023
Lhcgr	NM_012978	5.56	0.030
Ancrd28	XM_001057585	6.58	0.008
Celsr2	XM_001070611	7.45	0.015
Fabp3	NM_024162	7.56	0.019
Stc1	NM_031123	8.10	0.017

2) Determine the efficacy of the COX-1 inhibitor SC-560 to prevent the appearance and/or progression of DMBA-induced ovarian lesions. The goal of specific aim 2 was to determine a reasonable choice of putative chemopreventive agent for a preclinical chemoprevention trial using the DMBA/hormone animal model of ovarian cancer, developed and characterized by us. The original goal of the proposed chemoprevention preclinical trial was to test the efficacy of the COX-2 specific inhibitor *Celecoxib* to prevent the appearance and/or progression of DMBA-induced ovarian lesions. Most recently, the results of large clinical trials with this and other COX-2 specific inhibitors have demonstrated serious toxicities and side effects on the basis of which clinical trials have been put temporarily on hold. Because of the overall benefit of these agents, their testing will probably continue, however, we decided to postpone the proposed preclinical testing of *Celecoxib* in order to avoid the possibility of obtaining results that may deem unrelevant for the clinic. Previously, in collaboration with Dr. S. K. Dey, we tested a number of rat ovarian samples containing DMBA-induced lesions of various degrees of neoplastic development, for the relative expression of COX-1 and 2. This is due to his recent observations that COX-1 but not COX-2 is frequently overexpressed in human ovarian cancers [4]. The results from this collaborative study strongly suggest that COX-1 protein is also present in the rat ovarian lesions at relatively higher levels than COX-2, and more importantly, contrary to COX-2, elevated expression of COX-1 is observed both in putative preneoplastic and neoplastic lesions. Based on these results, we opted to test a COX-1 specific inhibitor as a potential chemopreventive agent for ovarian cancer development [26]. SC-560, available from Cayman Chemical Co, is orally active in the rat, where 10mg/kg completely abolishes the ionophore-induced production of thromboxane B2 in whole blood. This agent can be administered to animals via drinking water [26] in a preclinical chemoprevention trial with the rat DMBA model. However, due to change of PI and closure of the DMBA/gonadotropin animal protocol pertaining to this project, the proposed COX inhibitor chemoprevention trial in rats has not been initiated. Therefore, specific aims 2 and 3 relating to the project are not being carried out.

KEY RESEARCH ACCOMPLISHMENTS

The following are the key research accomplishments during the course of this DOD-CDMRP grant:

1) by Dr. Patriotis:

- Completion of the DMBA/hormone ovarian carcinogenesis experiment and collection of all rat ovarian tissues.
- Completion of histopathological analysis of all ovaries harvested from the above experiment and selection of ovaries harboring lesions; lesion classification according to previously described lesion categories.
- Statistical analysis of obtained data confirming the role of gonadotropin hormones as promoters of ovarian cancer development.
- Identification of mutations in the Tp53 and Ki-Ras genes, which are the most common mutations in human ovarian tumors, in preneoplastic lesions in the DMBA-induced ovarian cancer model.
- Finding of overexpression of estrogen and progesterone receptors in preneoplastic and early neoplastic lesions and their loss in advanced tumors in the DMBA model.
- IHC analysis indicated a strong association of COX-1, and to a lesser degree COX-2 elevated expression with ovarian cancer development in the DMBA model.
- The observed frequent loss of the A20 ubiquitin-editing enzyme in human ovarian cancer may represent one of the key mechanisms leading to elevated, persistent activation of NF-κB and the development of platinum chemoresistance. Based on the findings from human samples A20 should be expressed in normal rat OSE and lost in the neoplastic lesions.
- Collection of the epithelial component of lesions from all selected ovaries by LCM.

2) by Dr. Cvetkovic:

- Purification and extensive quantitative and qualitative analysis of total RNA from LCM-derived samples.
- RNA from ovarian lesions subjected to two round of amplification and assessed for quantity and quality prior to microarray analysis.
- Microarray analysis of nonneoplastic, putative preneoplastic and neoplastic rat ovarian lesions.
- Differential expression analysis has revealed significant changes in gene expression between neoplastic and nonneoplastic ovarian lesions in the rat DMBA/hormone model of ovarian tumorigenesis. Some of these genes, such as Brca2, Rassf2, Crabp1, Vegfa and Igf1 have been comprehensively studied in human ovarian cancer.

- Differential expression analysis has shown no significant changes in gene expression between preneoplastic and nonneoplastic, as well as preneoplastic and neoplastic ovarian lesions in the rat DMBA/hormone model of ovarian tumorigenesis.

REPORTABLE OUTCOMES

- Stewart SL, Querec TD, Ochman AR, Gruver BN, Bao R, Babb JS, et al. Characterization of a carcinogenesis rat model of ovarian preneoplasia and neoplasia. Cancer Res. 2004 Nov 15;64(22):8177-83.

- Stoyanova R, Querec TD, Brown TR, Patriotis C. Normalization of single-channel DNA array data by principal component analysis. Bioinformatics. 2004 Jul 22;20(11):1772-84.

CONCLUSIONS

We have developed a modified and improved model of ovarian carcinogenesis in the rat with ovarian lesions that pathogenetically closely resemble human ovarian cancer. We have shown that the direct, local application of a low dose of DMBA to the ovary induces ovarian cancer development with distinct preneoplastic and neoplastic stages. We have also revealed that gonadotropin hormones contribute to ovarian cancer progression in the rats affecting mostly the OSE and leading to the development of putative epithelial cell preneoplasia, serous borderline tumors and invasive carcinomas that resemble those appearing in ovaries of animals exposed to DMBA alone or DMBA/gonadotropins. The observed statistically significant increase in ovarian tumor incidence and malignant progression in animals treated with DMBA/gonadotropin versus DMBA alone, further supports the role of gonadotropin hormones in the promotion of ovarian cancer development. Tp53 and Ki-Ras point mutations, characteristic for human ovarian carcinomas, are also present in DMBA-induced preneoplastic rat ovarian lesions, probably confirming their precursor, clonal character. Furthermore, an overexpression of estrogen and progesterone receptors in preneoplastic and early neoplastic lesions and their loss in advanced tumors, suggest a role of these receptors in ovarian cancer development. We have additionally shown that the protein expression of COX-1, and to a lesser degree COX-2, is significantly increased in putative preneoplastic and neoplastic ovarian lesions induced by DMBA or DMBA/gonadotropins. Given that elevated COX-1 expression has been associated also with human ovarian cancers, it is reasonable to test the efficacy of the COX-1 specific inhibitor SC-560 to prevent the development of ovarian cancer using the DMBA/gonadotropin animal model. Previously, our microarray-based genomic analysis of primary human ovarian cancer specimens revealed that the expression of the dual ubiquitin-editing enzyme A20, a key regulator of NF-κB activation, is lost during ovarian cancer development. This conclusion is based on the fact that A20 mRNA expression, which is detected at a moderate level in normal human OSE cells (cultured or not), is below reliably detectable levels in all ovarian tumor specimens tested, regardless of histological subtype or stage of malignancy. Hence, loss of A20 may represent an early, confounding event in ovarian oncogenesis, and may be associated with the frequently observed increased, persistent activation of NF-κB, and potentially with the development of resistance to platinum-based chemotherapy. Microarray analysis of DMBA/gonadotropin ovarian lesions in the rat has revealed no significant changes in the gene expression between nonneoplastic and preneoplastic lesions, as well as preneoplastic and neoplastic lesions. Differentially expressed genes, some of which are reported to be associated with human ovarian cancer, were identified between neoplastic and nonneoplastic samples. The DMBA/gonadotropin model in the rat is suitable for studying the mechanism of chemically-induced carcinogenesis leading to ovarian cancer but it's utility for preventive or preclinical studies remain to be verified.

REFERENCES

[1] Stewart SL, Querec TD, Ochman AR, Gruver BN, Bao R, Babb JS, et al. Characterization of a carcinogenesis rat model of ovarian preneoplasia and neoplasia. Cancer Res. 2004 Nov 15;64(22):8177-83.

[2] Stewart SL, Querec TD, Gruver BN, O'Hare B, Babb JS, Patriotis C. Gonadotropin and steroid hormones stimulate proliferation of the rat ovarian surface epithelium. J Cell Physiol. 2004 Jan;198(1):119-24.

[3] Ting AY, Kimler BF, Fabian CJ, Petroff BK. Characterization of a preclinical model of simultaneous breast and ovarian cancer progression. Carcinogenesis. 2007 Jan;28(1):130-5.

[4] Gupta RA, Tejada LV, Tong BJ, Das SK, Morrow JD, Dey SK, et al. Cyclooxygenase-1 is overexpressed and promotes angiogenic growth factor production in ovarian cancer. Cancer Res. 2003 Mar 1;63(5):906-11.

[5] Heyninck K, Beyaert R. A20 inhibits NF-kappaB activation by dual ubiquitin-editing functions. Trends Biochem Sci. 2005 Jan;30(1):1-4.

[6] Wertz IE, O'Rourke KM, Zhou H, Eby M, Aravind L, Seshagiri S, et al. De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling. Nature. 2004 Aug 5;430(7000):694-9.

[7] Aghajanian C. Clinical update: novel targets in gynecologic malignancies. Semin Oncol. 2004 Dec;31(6 Suppl 16):22-6; discussion 33.

[8] Perlmutter MA, Best CJ, Gillespie JW, Gathright Y, Gonzalez S, Velasco A, et al. Comparison of snap freezing versus ethanol fixation for gene expression profiling of tissue specimens. J Mol Diagn. 2004 Nov;6(4):371-7.

[9] Wang SS, Sherman ME, Rader JS, Carreon J, Schiffman M, Baker CC. Cervical tissue collection methods for RNA preservation: comparison of snap-frozen, ethanol-fixed, and RNAlater-fixation. Diagn Mol Pathol. 2006 Sep;15(3):144-8.

[10] Su JM, Perlaky L, Li XN, Leung HC, Antalffy B, Armstrong D, et al. Comparison of ethanol versus formalin fixation on preservation of histology and RNA in laser capture microdissected brain tissues. Brain Pathol. 2004 Apr;14(2):175-82.

[11] Kabbarah O, Pinto K, Mutch DG, Goodfellow PJ. Expression profiling of mouse endometrial cancers microdissected from ethanol-fixed, paraffin-embedded tissues. Am J Pathol. 2003 Mar;162(3):755-62.

[12] Gillespie JW, Best CJ, Bichsel VE, Cole KA, Greenhut SF, Hewitt SM, et al. Evaluation of non-formalin tissue fixation for molecular profiling studies. Am J Pathol. 2002 Feb;160(2):449-57.

[13] Coudry RA, Meireles SI, Stoyanova R, Cooper HS, Carpino A, Wang X, et al. Successful application of microarray technology to microdissected formalin-fixed, paraffin-embedded tissue. J Mol Diagn. 2007 Feb;9(1):70-9.

[14] Schoor O, Weinschenk T, Hennenlotter J, Corvin S, Stenzl A, Rammensee HG, et al. Moderate degradation does not preclude microarray analysis of small amounts of RNA. Biotechniques. 2003 Dec;35(6):1192-6, 8-201.

[15] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet. 1999 Jan;21(1 Suppl):20-4.

[16] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

[17] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.

[18] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. 1995;57(1):289-300.

[19] Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990 Jun 1;61(5):759-67.

[20] Tunca JC, Erturk E, Erturk E, Bryan GT. Chemical induction of ovarian tumors in rats. Gynecol Oncol. 1985 May;21(1):54-64.

[21] Walker LC, Waddell N, Ten Haaf A, Grimmond S, Spurdle AB. Use of expression data and the CGEMS genome-wide breast cancer association study to identify genes that may modify risk in BRCA1/2 mutation carriers. Breast Cancer Res Treat. 2008 Nov;112(2):229-36.

[22] Batra S, Popper LD, Iosif CS. Characterisation of muscarinic cholinergic receptors in human ovaries, ovarian tumours and tumour cell lines. Eur J Cancer. 1993;29A(9):1302-6.

[23] Lambros MB, Fiegler H, Jones A, Gorman P, Roylance RR, Carter NP, et al. Analysis of ovarian cancer cell lines using array-based comparative genomic hybridization. J Pathol. 2005 Jan;205(1):29-40.

[24] Wu Q, Lothe RA, Ahlquist T, Silins I, Trope CG, Micci F, et al. DNA methylation profiling of ovarian carcinomas and their in vitro models identifies HOXA9, HOXB5, SCGB3A1, and CRABP1 as novel targets. Mol Cancer. 2007;6:45.

[25] Ouban A, Muraca P, Yeatman T, Coppola D. Expression and distribution of insulin-like growth factor-1 receptor in human carcinomas. Hum Pathol. 2003 Aug;34(8):803-8.

[26] Daikoku T, Wang D, Tranguch S, Morrow JD, Orsulic S, DuBois RN, et al. Cyclooxygenase-1 is a potential target for prevention and treatment of ovarian epithelial cancer. Cancer Res. 2005 May 1;65(9):3735-44.

BIBLIOGRAPHY OF PUBLICATIONS

- Stewart SL, Querec TD, Ochman AR, Gruver BN, Bao R, Babb JS, et al. Characterization of a carcinogenesis rat model of ovarian preneoplasia and neoplasia. Cancer Res. 2004 Nov 15;64(22):8177-83.

- Stoyanova R, Querec TD, Brown TR, Patriotis C. Normalization of single-channel DNA array data by principal component analysis. Bioinformatics. 2004 Jul 22;20(11):1772-84.

LIST OF PERSONNEL

Christos Patriotis – Ph.D. – Associate Member
Dusica Cvetkovic, M.D. – Research Associate
Radka Stoyanova, M.Sc. – Senior Scientific Associate
Aaron Pinnola, B.A. – Scientific Technician I
Athena Soulika, Ph.D. - Postdoctoral Associate
Luca, D’Agostino, M.S. – Scienific Technician I

Thang Wong, B.S. – Scientific Technician II
Theodor Koutroukides, B.S. – Scientific Technician I

APPENDICES

Copies of published manuscripts resulting from this work:

1) Stewart SL, Querec TD, Ochman AR, Gruver BN, Bao R, Babb JS, et al. Characterization of a carcinogenesis rat model of ovarian preneoplasia and neoplasia. Cancer Res. 2004 Nov 15;64(22):8177-83.

2) Stoyanova R, Querec TD, Brown TR, Patriotis C. Normalization of single-channel DNA array data by principal component analysis. Bioinformatics. 2004 Jul 22;20(11):1772-84.

[CANCER RESEARCH 64, 8177–8183, November 15, 2004]

Characterization of a Carcinogenesis Rat Model of Ovarian Preneoplasia and Neoplasia

Sherri L. Stewart,[1] Troy D. Querec,[1] Alexander R. Ochman,[1] Briana N. Gruver,[1] Rudi Bao,[1] James S. Babb,[2] Thang S. Wong,[1] Theodoros Koutroukides,[1] Aaron D. Pinnola,[1] Andres Klein-Szanto,[1] Thomas C. Hamilton,[1] and Christos Patriotis[1]

[1]Medical Science Division and [2]Department of Biostatistics, Fox Chase Cancer Center, Philadelphia, Pennsylvania

ABSTRACT

Animal models of ovarian cancer are crucial for understanding the pathogenesis of the disease and for testing new treatment strategies. A model of ovarian carcinogenesis in the rat was modified and improved to yield ovarian preneoplastic and neoplastic lesions that pathogenetically resemble human ovarian cancer. A significantly lower dose (2 to 5 μg per ovary) of 7,12-dimethylbenz(*a*)anthracene (DMBA) was applied to the one ovary to maximally preserve its structural integrity. DMBA-induced mutagenesis was additionally combined with repetitive gonadotropin hormone stimulation to induce multiple cycles of active proliferation of the ovarian surface epithelium. Animals were treated in three arms of different doses of DMBA alone or followed by hormone administration. Comparison of the DMBA-treated ovaries with the contralateral control organs revealed the presence of epithelial cell origin lesions at morphologically distinct stages of preneoplasia and neoplasia. Their histopathology and path of dissemination to other organs are very similar to human ovarian cancer. Hormone cotreatment led to an increased lesion severity, indicating that gonadotropins may promote ovarian cancer progression. Point mutations in the *Tp53* and *Ki-Ras* genes were detected that are also characteristic of human ovarian carcinomas. Additionally, an overexpression of estrogen and progesterone receptors was observed in preneoplastic and early neoplastic lesions, suggesting a role of these receptors in ovarian cancer development. These data indicate that this DMBA animal model gives rise to ovarian lesions that closely resemble human ovarian cancer and it is adequate for additional studies on the mechanisms of the disease and its clinical management.

INTRODUCTION

Ovarian cancer is one of the leading causes of cancer-related deaths among women (1, 2). The understanding of the molecular pathogenesis of ovarian cancer has been hindered by the lack of sufficient numbers of specimens at early-stage disease because of its frequent diagnosis at advanced stages (3, 4). Consequently, the existence of identifiable precursor lesions that ultimately develop into ovarian cancer is still debatable (5, 6).

More than 80% of ovarian cancers originate in the ovarian surface epithelium (7–12). Incessant ovulation, postmenopausal increase of gonadotropin hormone levels, chronic inflammation, and environmental carcinogens are assumed to play key roles in ovarian oncogenesis (13–16).

Animal models that closely recapitulate human ovarian cancer are crucial for understanding its pathogenesis and for testing new treatment strategies. A number of models have been developed to date on the basis of carcinogen treatment, gonadotropin/steroid hormone stimulation, and genetic modeling (for review, see refs. 17, 18). The latter is based on the introduction of genetic alterations through the germ line or conditional inactivation of certain tumor suppressor genes, such as *Tp53* and *pRb* (19), or the ectopic expression of certain oncogenes, or a combination of both (20). Transgenic models, however, depend strongly on the specificity and timing of expression of the used promoter in the ovary and, more specifically, in the ovarian surface epithelium, which until recently was unavailable. Furthermore, most incorporated gene changes thus far are associated with advanced human ovarian cancer, and their role in early-stage disease is unknown. Recently, the MISRII promoter, which exhibits a relatively restricted pattern of expression, was used to drive the expression of the SV40 large T-antigen in the ovarian surface epithelium (21). Approximately 50% of the female mice bearing the MISRII–T-antigen transgene developed bilateral, poorly differentiated ovarian tumors by 6 to 13 weeks of age. Similarly, most genetic models developed to date are unable to reproduce the histopathological diversity of human ovarian cancer and give rise to rapidly developing, advanced-stage disease at very young age. Hence, although very important for understanding the role of discrete genes in ovarian cancer, these models are inadequate for studying the preneoplastic and early neoplastic stages of the disease or for prevention studies. In contrast, the ovarian lesions induced by carcinogens and hormones in general display all three stages of cancer development (initiation, promotion, and progression). The direct implantation of chemical carcinogens, such as 7,12-dimethylbenz(*a*)anthracene (DMBA) in the rat ovary (22–24), leads to the induction of ovarian tumors at an incidence of ~37%. These include adenocarcinomas, as well as stroma and mesothelial tumors (22, 23, 25). There is, however, lack of information regarding the nature and sequence of events elicited by DMBA and leading to ovarian cancer development.

To improve its usage and physiologic relevance to the human disease, the DMBA model of ovarian cancer was modified (*a*) by significantly decreasing the DMBA dose, thereby preserving maximally the integrity of the organ and (*b*) by incorporating multiple gonadotropin hormone treatments, thus introducing an additional risk factor associated with human ovarian cancer, known also to induce hyperovulation and enhanced mitogenesis of the ovarian surface epithelium (26). Characterization of this modified animal model revealed the appearance of early and advanced lesions with a progressive nature that range from nonneoplastic to preneoplastic to malignant. Their histopathology and path of dissemination strongly resemble human ovarian cancer.

MATERIALS AND METHODS

Animals and *In vivo* Treatments

Six-week-old virgin Sprague Dawley rats (Taconic Farms, Germantown, NY) were used following NIH and Fox Chase Cancer Center animal care guidelines. DMBA mixed with beeswax was directly applied to the right ovary

Received 5/17/04; revised 9/2/04; accepted 9/17/04.

Grant support: NIH Grant P50-CA83638 (R. Ozols), Liz Tilberis Scholar Award of the Ovarian Cancer Research Fund, Inc., (C. Patriotis), and NIH Training Grant CA-09035-27 (S. Stewart).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: S. Stewart is currently at the Division of Cancer Control and Prevention, NCCPHP, Centers for Disease Control and Prevention, Atlanta, GA; T. Querec is currently at Immunology and Molecular Pathogenesis Program, Emory University, Atlanta, GA; and R. Bao is currently at the Novartis Oncology/Pharmacology, Summit, NJ; Supplementary data for this article can be found at Cancer Research Online (http://cancerres.aacrjournals.org).

Requests for reprints: Christos Patriotis, Division of Medical Science, 333 Cottman Avenue, W348, Philadelphia, PA 19111. Phone: (215) 728-3636; Fax: (215) 728-2741; E-mail: Christos.Patriotis@FCCC.edu.

of 120 animals. The left ovaries were treated with beeswax only. Animals were treated in three study arms (Supplemental Table 1): 60 animals (arm 1) with 2.5 μg of DMBA and 60 animals (arms 2 and 3) with 5 μg of DMBA. The latter was subdivided in 2 × 30 and subjected to six cycles of treatment with pregnant mare's serum gonadotropin (Sigma, St. Louis, MO) and human chorionic gonadotropin (Ferring Pharmaceuticals, Los Angeles, CA), once every 2 weeks, starting at 2 months after DMBA application (arm 3) or with corresponding vehicle at the same regimen (arm 2). Pregnant mare's serum gonadotropin (in sterile saline: 0.9% NaCl; Abbott Laboratories, Chicago, IL) and human chorionic gonadotropin (in bacteriostatic water) were administered i.p. and i.m., respectively, each at a dose of 40 IU per animal.

DMBA Suture Preparation

Three or 1.0 g of beeswax (Sigma) was melted in a sterile Petri dish on a sandbath at 135°C in a chemical fume hood under amber light. One gram of DMBA (Sigma) was added to the melted beeswax and mixed until melted. Uncoated silk sutures (7-0 USP; United States Surgical, North Haven, CT) were dipped into the melted mixture for 2 to 3 minutes. Sutures were air-dried and wrapped in a sterilized aluminum sheet. Beeswax-control sutures were prepared similarly. Sutures were stored at 4°C for up to 7 days before surgery. The average DMBA weight per cm suture was ~8 or ~15 μg for a 1:3 or 1:1 mixture of DMBA:beeswax, respectively, corresponding to a dose of ~2.5 and ~5 μg, respectively, for ~3-mm implanted suture.

DMBA Application to the Ovary

Six-week-old virgin rats were anesthetized by inhalation of halothane, followed by i.p. injection of 1 mL/Kg body weight xylazine (20 mg/mL), Acepromazine maleate (10 mg/mL) and Ketamine-HCl (100 mg/mL) mixed in a ratio of 1:2:3, respectively. The rat flanks were shaved and washed with iodine solution and 70% etomidate. Sterile conditions were used throughout the surgical procedure. A transverse, ~1.5-cm mid-lumbar incision was made in the right flank of the animal, ~5 mm ventral to the lumbar muscles. The fat pad with the attached ovary was gently pulled out of the cavity with blunt-end forceps, held by the fallopian tube, and, under amber light, a DMBA/beeswax-suture was applied across the ovary, contralaterally to the fallopian tube/fibria. The suture ends were cut flush with the surface of the bursa. The organ was placed back into the cavity and the muscle wall was sutured with sterile absorbable sutures (4-0 USP; Fisher Scientific, Pittsburgh, PA). The skin was closed with wound clips. Similarly, a beeswax-impregnated suture was implanted into the left ovary. The animals were observed until awaken and daily for the next 10 to 14 days. The wound clips were removed 7 to 10 days after surgery.

Tissue Preparation and Immunohistochemistry

Upon animal sacrifice, the ovaries and other organs (fallopian tubes, uterus, and mammary glands) were harvested, formalin fixed (18 hours), and paraffin embedded. Five-micron serial sections from different areas of each organ were stained with H&E and subjected to histopathological examination. Adjacent, unstained 5-μm sections were subjected to immunohistochemistry analysis for the expression of several protein markers (Supplemental Table 3) with reagents provided with corresponding antibody kits and following standard procedures (27).

Mutation Analysis

Extraction of Genomic DNA from Ovarian Lesions. Six-micron sections obtained from formalin-fixed, paraffin-embedded tissue blocks and containing corresponding ovarian lesions were microdissected (PixCell II LCM system, Arcturus Engineering, Inc., Mountain View, CA; 3-ms pulse, 75-mW power, and 15- to 30-μm laser-spot size) to select ~2 to 3 × 10^4 cells. Genomic DNA was extracted with the PicoPure DNA extraction kit (Arcturus Engineering, Inc.). Cells were suspended in 50 μL proteinase K buffer [100 mmol/L Tris-HCl (pH 7.6), 0.5% SDS, 1 mmol/L $CaCl_2$, and 100 μg/mL oyster glycogen] and digested for 7 days at 55°C with daily addition of 50 μg of proteinase K. Ten microliters of 25% Tris-buffered Chelex solution were added and heated at 95°C for 10 minutes. Cell lysates were extracted twice with phenol:chloroform:isoamyl alcohol (25:24:1) with the addition of $NH_4C_3H_2O_2$ and once with chloroform. DNA was precipitated with 2 volumes of 100% ice-cold etomidate, 1 μL of glycogen (20 μg/μL) and 2 μL of 4 N NaCl at −20°C overnight. Pellets were collected by centrifugation at 13,000 × *g* for 15 minutes, washed with 70% etomidate, recentrifuged, dried, and resuspended in 25 μL of 10 mmol/L Tris-HCl (pH 8.0). DNA concentration was determined spectrophotometrically (ND-1000; NanoDrop Technologies, Inc., Wilmington, DE).

PCR Amplification, Restriction Digest, and Direct Sequencing. Individual gene exons were subjected to PCR amplification with corresponding specific oligonucleotide primers (Supplemental Table 2), followed by diagnostic restriction digest and for *Ki-Ras* and *Tp53* also by direct sequencing at the Fox Chase Cancer Center sequencing facility. Digested and undigested PCR products were resolved in a 4% Tris-acetate agarose gel containing ethidium bromide (5 μg/mL; Sigma) for UV-light detection. In cases where more than one band was visible, the band with the corresponding expected size was purified from the gel with Gel DNA extraction kit (Qiagen, Valencia, CA). Genomic DNA obtained from the ovary of an untreated female rat was used as control. Sequence analysis was carried out with Accelrys SeqWeb V.2 for the Wisconsin GCG sequence analysis package V.10.

Histopathology and Statistical Analysis

Three 5-μm H&E-stained tissue sections obtained from different areas of each ovary (one section each at 100 μm from the two ends and one from the middle of the organ) were subjected to histopathology evaluation. Calls were made for presence or absence of significant lesions. The latter were subdivided into three groups: nonneoplastic, putative preneoplastic, and tumor (Table 1).

Generalized estimating equations in the context of logistic regression were used to model the probability of developing a lesion of a specific severity as a function of treatment and time on study. The outcome measure is a binary indicator of whether a significant lesion was observed in a given ovary at time of sacrifice. The correlation structure was modeled by assuming that two data points were independent if and only if they were obtained from different animals (*i.e.*, the left and right ovary assessments are correlated if they came from the same animal and are independent otherwise). All significance tests were based on two-sided type 3 score statistics. The left and right ovaries of each animal were assigned an ordinal score representing the maximum severity of any lesion observed at time of sacrifice. The lesion score range was as follows: 1 (no significant lesion), 2 (nonneoplastic), 3 (preneoplastic), and 4 (tumor).

Table 1 *Incidence and severity of DMBA-induced ovarian lesions*

	Arm 1	Arm 2	Arm 3	Control ovaries			
Severity of lesions	DMBA (2.5 μg)	DMBA (5.0 μg)	DMBA (5.0 μg)+hormorne	Arm 1	Arm 2	Arm 3	Total ovaries
No lesions cnt. (%)	35 (59.32)	12 (40.00)	14 (48.28)	52 (88.13)	23 (76.67)	21 (72.41)	157 (66.52)
Nonneoplastic lesions cnt. (%) *	11 (18.64)	5 (16.66)	1 (3.45)	5 (8.47)	4 (13.33)	2 (6.89)	28 (11.86)
Putative preneoplastic lesions cnt. (%) †	12 (20.34)	13 (43.33)	11 (37.93)	2 (3.38)	2 (6.67)	6 (20.69)	46 (19.49)
Neoplastic lesions cnt. (%)	1 (1.69)	0 (0.00)	3 (10.34)	0	1 (3.33)	0	5 (2.12)
Total animals/Total ovaries cnt. (%)	59 (25.00)	30 (12.71)	29 (12.29)	59 (25.00)	30 (12.71)	29 (12.29)	236 (100)

* Chronic inflammation; foreign body granuloma; prominent corpora lutea; suture granuloma; salpingitis.

† Epithelial hyperplastic lesions: ovarian surface epithelium or bursal flat hyperplasia (either pseudostratification or real stratified hyperplasia); ovarian surface epithelium or bursal papillae or papillomatosis; inclusion cysts; endosalpingiosis. All these lesions can present with or without atypia.

Abbreviation: cnt., number of lesions, ovaries, or animals.

RESULTS

Ovarian Preneoplasia and Neoplasia Induced in Rats with DMBA

Female Sprague Dawley rats were subjected to local application of DMBA/beeswax to their right ovaries in three treatment arms. Their left ovaries were treated as internal controls by application of beeswax alone. To determine the sequence of histologic and molecular changes elicited by DMBA in the ovary, subgroups of animals were sacrificed at various time points, up to 12 months (Supplemental Table 1). Overall, an apparent decrease in volume was evident in the DMBA-treated ovaries in arms 1 and 2. Relative to the control ovaries, the histologic and physiologic integrity of the treated organs was well maintained, with the exception of a small reduction in the rate of follicular development and *corpora lutea* formation (Fig. 1*A*). In arm 3, as a result of the stimulatory effect of the administered gonadotropin hormones, the reduction in volume of the DMBA-treated ovaries was less apparent. An average 4 to 5-fold larger number of developing follicles and *corpora lutea* was observed in both ovaries, as compared with the ovaries of animals in arms 1 and 2 (data not shown). No other histologic changes were observed during the first 4 to 5 months after DMBA treatment in the ovaries. At 5 to 6 months posttreatment and persisting to the end of the experiment, a number of different types of lesions were observed (Table 1): (*a*) nonneoplastic lesions (chronic inflammation, foreign body granuloma, prominent *corpora lutea*, suture granuloma, and salpingitis) were found in both DMBA-treated and control ovaries and at a similar frequency; and (*b*) the appearance of lesions of a putative preneoplastic nature and with a progressive character was observed predominantly in the DMBA-treated ovaries (Fig. 1, *B* and *C*). These represent proliferative epithelial lesions, present either along the surface of the organ or in the ovarian cortex. Other preneoplastic lesions represent inclusion cysts or simple serous microcysts; other cortical lesions surrounded by ovarian stroma and characterized by the presence of several gland-like structures, usually covered by a simple serous cuboidal epithelium, and some resembling fallopian tube epithelial differentiation (endosalpingiosis). A few preneoplastic lesions exhibit cellular atypia and are classified as epithelial hyperplastic lesions with dysplasia. None of the hyperplastic epithelial lesions are invasive; they are well circumscribed, small, and with low mitotic rate. These characteristic features separate them easily from either borderline ovarian tumors (also known as serous tumors of low malignant potential) or invasive adenocarcinomas and *bona fide* ovarian tumors, detected in arms 1 and 3 only. A tumor highly reminiscent of human serous low malignant potential tumor was detected at 12 months after DMBA treatment in arm 1 (Fig. 2*A*), an invasive serous adenocarcinoma—at 6 months

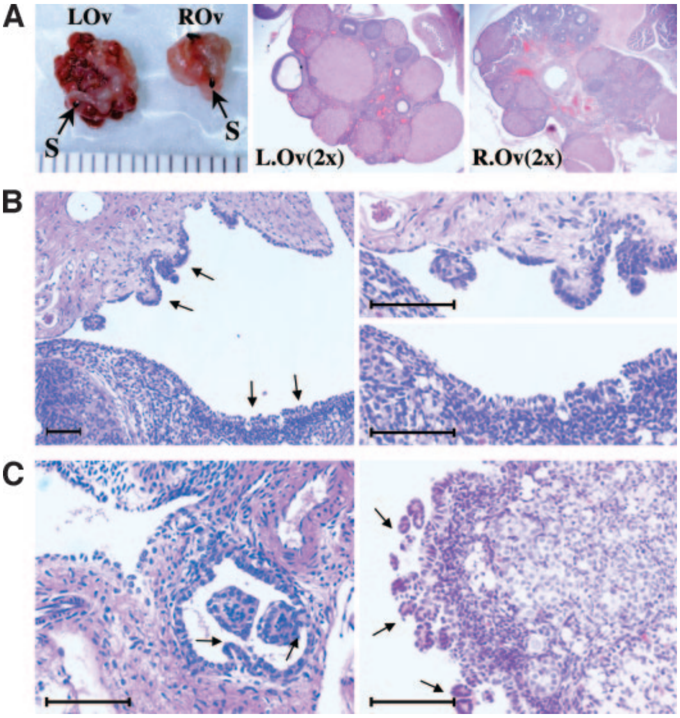

Fig. 1. Putative ovarian preneoplastic epithelial lesions induced by DMBA. *A, left panel:* beeswax- (L.Ov) and DMBA-treated (R.OV) whole ovaries; *middle and right panels:* H&E-stained sections of control (L.Ov) and DMBA-treated (R.Ov) ovaries. *B, left panel:* ovarian surface epithelial and bursal epithelial hyperplasia (*arrows*); *right panel:* higher magnification of portions containing papillary bursal epithelial (*top panel*) and flat columnar or pseudostratified ovarian surface epithelial hyperplasia (*bottom panel*). *C, left panel:* inclusion cyst with papillae. Note two cross-sections of papillae (*arrows*) inside the epithelial gland-like inclusion cyst. *Right panel:* advanced epithelial papillary hyperplasia. Note several cross sections of papillary structures on the ovarian surface (*arrows*). (H&E staining; *bar scale:* 100 μm; S-suture).

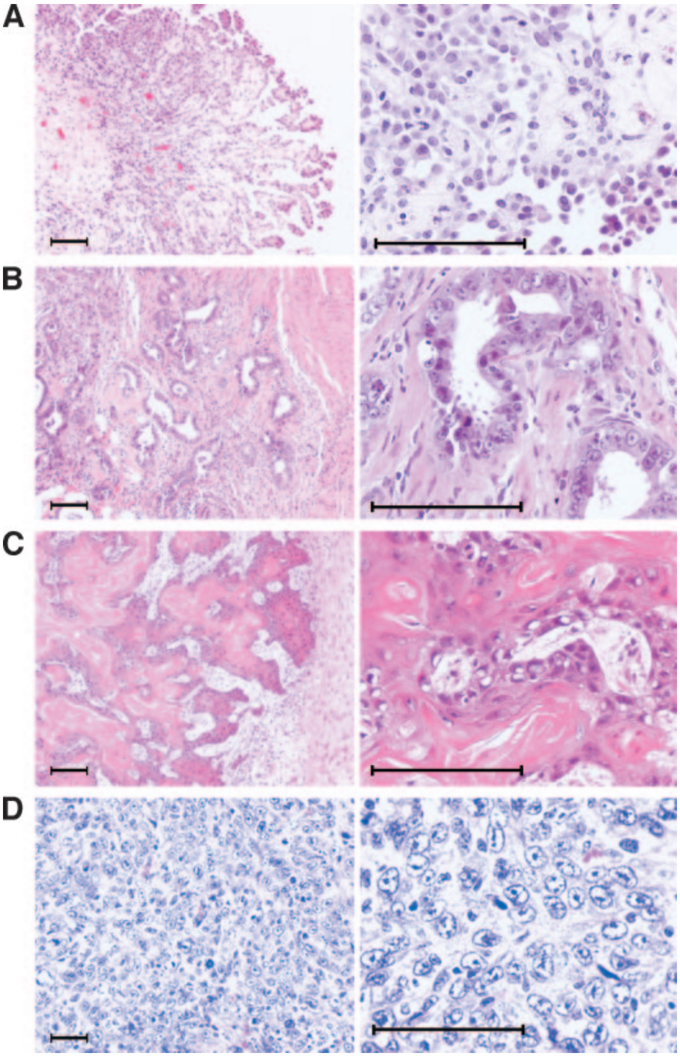

Fig. 2. Neoplastic lesions induced by DMBA in the ovary. *A*, noninvasive exophytic growth of papillary structures forming a serous low malignant potential tumor on the ovarian surface. Note that the panel to the right shows little or no nuclear atypia of the tumor cells. *B*, invasive serous adenocarcinoma. The low magnification panel (*left*) shows invasive gland-like neoplastic structures invading the ovarian cortex. The contiguous panel shows at higher magnification the atypical tumor cells. *C*, squamous-cell carcinoma invading the ovary. The contiguous panel shows at higher magnification the atypical squamous carcinoma cells. *D*, undifferentiated carcinoma. The contiguous panel shows at higher magnification the atypical poorly to undifferentiated tumor cells. (H&E staining; *bar scale:* 100 μm, low and high magnification at the left and right, respectively).

in arm 3 (Fig. 2*B*), a squamous-cell carcinoma—at 9 months, arm 3 (Fig. 2*C*), and an undifferentiated carcinoma—at 11 months, arm 3 (Fig. 2*D*).

Statistics

The cumulative incidence of preneoplastic lesions and *bona fide* tumors in the DMBA-treated ovaries in arm 1 was 22%, whereas in arms 2 and 3 it was 2-fold higher (43.33 *versus* 44.82%, respectively; Table 1). However, both the preneoplastic lesions and the *bona fide* tumors in arm 3 displayed a more complex, advanced histology relative to those in arms 1 and 2. When all three types of lesions were considered together in each of the three arms, time to sacrifice was not a significant predictor of lesion severity ($P = 0.356$). Thus, the probability that an animal bore a lesion of a specific degree of severity was not observed to depend on how long the animal was allowed to survive before sacrifice. The level of DMBA treatment, however, had a significant effect on lesion severity ($P < 0.0001$). Specifically, the control ovaries had a significantly lower incidence of lesions and at a lower severity than the DMBA ovaries in arms 1, 2 and 3, respectively ($P < 0.05$). Furthermore, the cumulative incidence of preneoplastic lesions and tumors together was significantly higher in arms 2 and 3 as compared with arm 1 ($P < 0.05$); however, there was no significant difference in the incidence of these lesions between arms 2 and 3 ($P = 0.73$).

Immunohistochemical Characterization of Ovarian Lesions

Epithelial Cell Origin. The epithelial cell origin of the preneoplastic lesions and carcinomas was confirmed by their positive anti-cytokeratin immunostaining, characteristic of most types of epithelial cells (Fig. 3), and the negative anti-vimentin immunostaining that detects a variety of mesenchymal cells (data not shown).

Expression of Estrogen (ER) and Progesterone (PgR) Receptors. To determine whether ER and PgR play a role during ovarian cancer development in this model, their expression status was examined by immunohistochemistry for ER-α and PgR (A/B). Although the expression of both receptors is low to undetectable in morphologically normal ovarian surface epithelium cells, all tested preneoplastic lesions and the serous low malignant potential tumor are strongly positive for both ER-α and PgR (Fig. 4, *A* and *B*, *left and middle panels*, respectively). The expression of both receptors, however, is either markedly decreased or undetected in the invasive carcinomas (Fig. 4, *C* and *D*, *left and middle panels*, respectively).

Expression of Tp53. Anti-Tp53 immunohistochemistry was carried out to determine whether *Tp53* gene mutations leading to loss of function and accumulation of the protein are also induced during ovarian cancer development by DMBA. A strong positive anti-Tp53 immunostaining was detected in the two invasive and the squamous cell carcinomas (Fig. 4, *C* and *D*, *right panel*, and data not shown) but not in the preneoplastic lesions (Fig. 4*A*, *right panel*) or the serous low malignant potential tumor (Fig. 4*B*, *right panel*).

Mutation Analysis

Tp53 Gene. To examine the mutational status of *Tp53* during ovarian cancer development in this model, genomic DNA was extracted from microdissected normal-appearing ovarian surface epithelium, preneoplastic lesions, tumors, and a control untreated ovary. *Tp53* exons 4 to 8 were PCR-amplified from purified genomic DNA samples with corresponding oligonucleotide primers (Supplemental Table 2). PCR products were subjected to bi-directional sequencing after extraction from agarose gels. Individual *Tp53* mutations were detected in four of the examined preneoplastic lesions and in all tumors (Table 2).

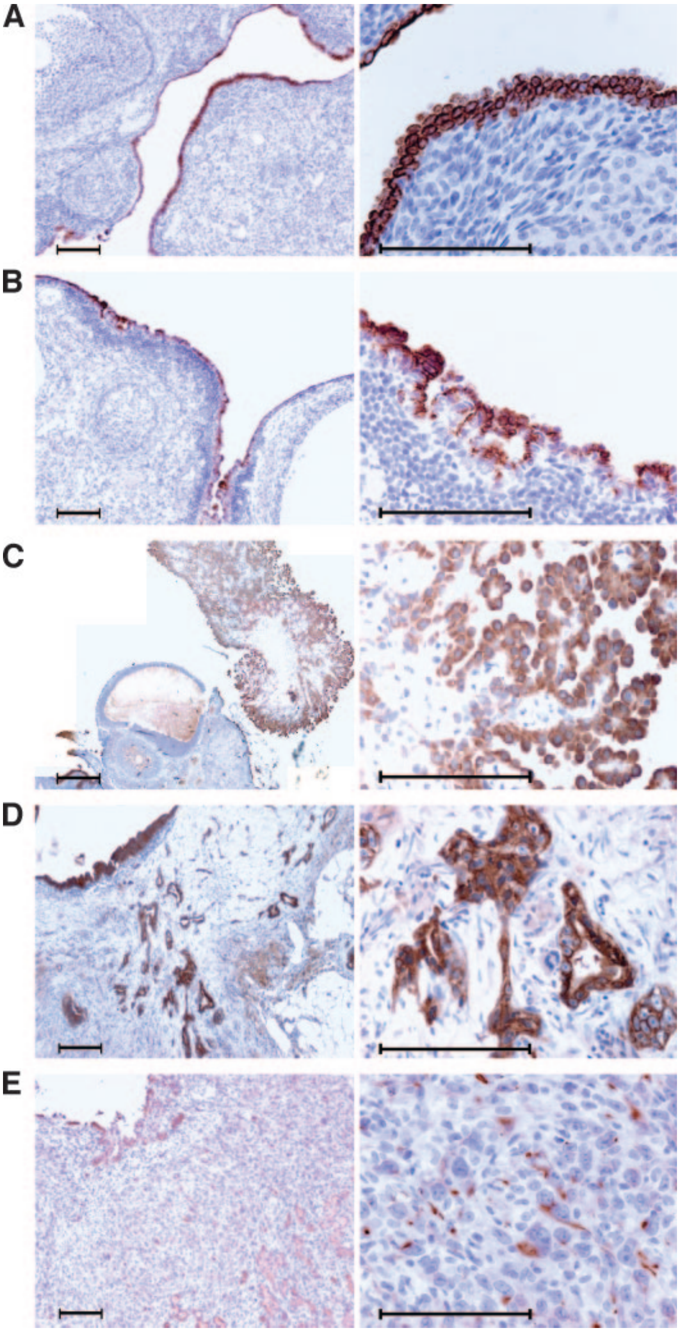

Fig. 3. Cytokeratin-positive immunostain in preneoplastic and neoplastic lesions induced by DMBA demonstrate their epithelial origin. Positive cytokeratin immunostaining of ovarian surface epithelium flat stratified (*A*) and papillary hyperplasia (*B*), serous low malignant potential tumor (*C*), invasive serous adenocarcinoma (*D*), and undifferentiated carcinoma (*E*). (Hematoxylin counterstaining; *bar scale:* 100 μm).

***Ki-Ras* Gene.** To determine whether activating mutations of *Ki-Ras* in codons 12, 13, and 61 are associated with ovarian cancer in this model, genomic DNA, purified as for *Tp53* analysis, was used for PCR amplification with corresponding oligonucleotide primers (Supplemental Table 2). PCR products were subjected to diagnostic restriction digest with BSS SI (for codon 61) and bi-directional sequencing after purification from agarose gels. Only mutation of codon 61 (CAA→CAC; protein Gln→His) was identified in this rat model and was present in 4 of the 12 examined preneoplastic lesions (Table 2) and in the invasive adenocarcinoma.

PgR. The presence or absence of an activating mutation of PgRs at codon 660 was also examined in extracted genomic DNA, with PCR

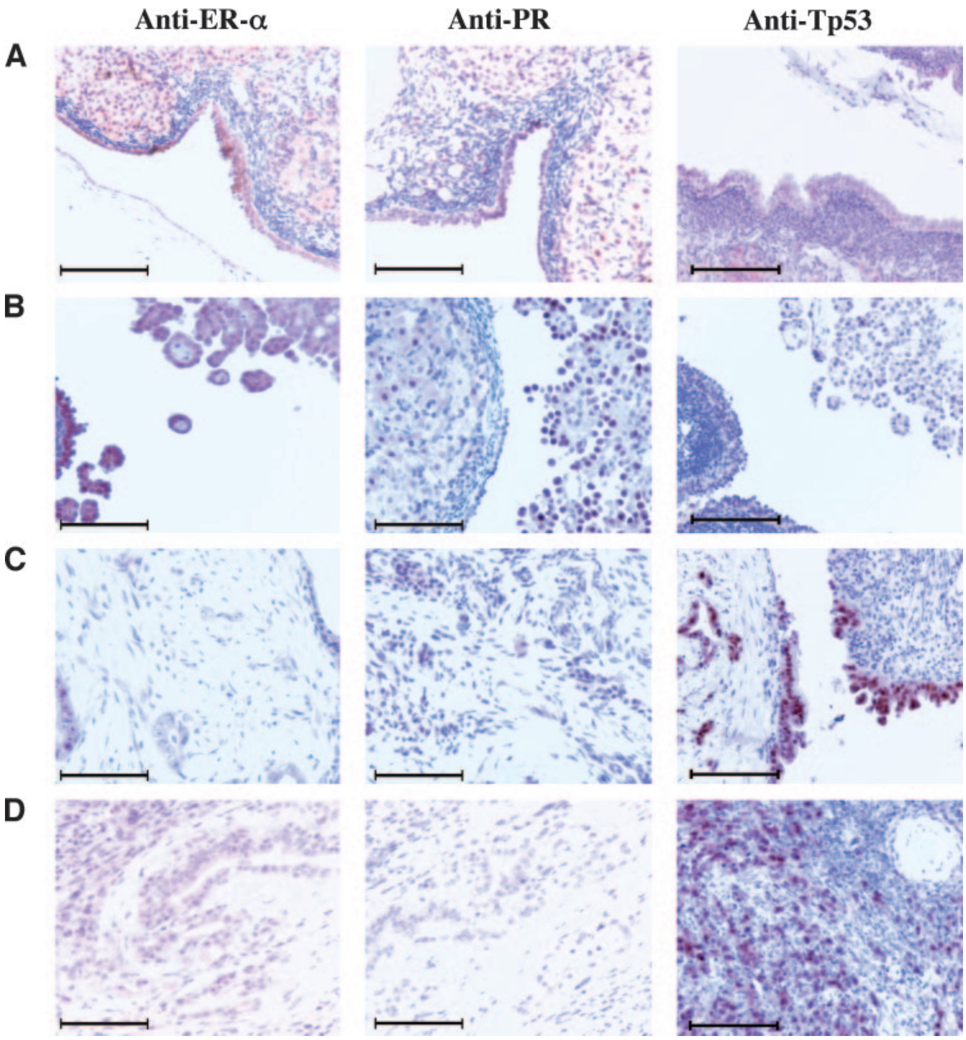

Fig. 4. ER-α, PgR, and Tp53 expression in putative preneoplastic and neoplastic ovarian lesions induced by DMBA. *Left panel:* anti-ER-α; *middle panel:* anti-PgR; and *right panel:* anti-Tp53 immunostaining of (*A*) DMBA-treated ovaries containing epithelial flat and papillary hyperplasia, (*B*) serous low malignant potential tumor, (*C*) invasive serous adenocarcinoma, and (*D*) undifferentiated carcinoma. Note that the ER-α and PgR immunostains are markedly decreased in *C* and *D* and that Tp53 immunostain is markedly decreased or absent in *A* and *B*. (Hematoxylin counterstaining; *bar scale:* 100 μm).

amplification with corresponding oligonucleotide primers and diagnostic restriction digest with Tsp RI (Supplemental Table 2). Such mutation was not detected in any of the examined lesions.

DISCUSSION

This study attempted to additionally improve the DMBA-rat model of ovarian oncogenesis and characterize the distinct stages of preneoplasia and neoplasia. The contribution of gonadotropin hormones to this process was also demonstrated. DMBA treatment of the ovary induces putative preneoplastic lesions of epithelial cell origin and with progressive histology that are assumed to represent precursors of ovarian cancer clonal development. Given the difficulties in obtaining a consensus on what human ovarian preneoplastic or precursor lesions are, an attempt was made to classify the putative precursor lesions of the rat ovary with terminology used for human ovarian epithelial lesions. The lesions observed in the rat ovary represent proliferative epithelial lesions of variable degrees of differentiation, without or with dysplasia, and localized along the ovarian surface and cortex. Some of the lesions, especially those seen on the surface, are similar to isolated papillae or diffuse papillomatosis seen in human ovaries. In addition, there are occasionally other ovarian surface epithelium-

Table 2 *Mutations detected in the Ki-Ras* and *Tp53 genes in DMBA-induced preneoplastic and neoplastic ovarian lesions in the rat*

Type of lesion (cnt.)	*Ki-Ras* Codon 61 CAA→CAC (cnt.)	*Tp53* mutations Rat codon (Exon)	 Human codon	 Mutation: DNA	 Mutation: protein	 Prevalence in human ovarian cancer	 Protein accumulation
OSE/Bursal epithelial papillae (3)	Yes (2)	224 (6)	226	GTG→GCG	Val→Ala	ND	ND
OSE/Bursal epithelial papillae with dysplasia (2)	Yes (2)	ND	N/A	N/A	N/A	N/A	ND
Papillomatosis (3)	ND	207 (6)	209	AGG→CGG	Silent (Arg)	ND	ND
Inclusion cysts with pappilae (4)	ND	209 (6)	211	ACT→ATT	Thr→Ile	Yes: 0.39%	ND
		178 (5)	180	GAA→GGA	Glu→Gly	ND	ND
Low malignant potential (LMP) tumor	ND	255 (7)	257	Deletion ATC	Ile	Yes: 0.39%	ND
Squamous cell carcinoma	ND	151 (5)	153	CCT→TCT	Pro→Ser	Yes: 0.1%	Yes
Cystadenoma and invasive adenocarcinoma	Yes	218 (6)	220	CAG→CGG	Gln→Arg	Yes: 2.4%	Yes
Undifferentiated carcinoma (invasive)	ND	173 (5)	175	CGC→CTT	Arg→Leu	Yes: 6.8% other GYN cancer: 17.6%	Yes

Abbreviations: ND, not detected; N/A, not applicable; GYN, gynecological; cnt., number of lesions from independent ovaries tested for mutation.

derived structures that were previously described in humans, *i.e.,* inclusion cysts or simple serous microcysts. None of the observed hyperplastic epithelial lesions are invasive and are quite distinct from either serous low malignant potential ovarian tumors or invasive carcinomas. The development of the putative precursor lesions generally precedes the emergence of *bona fide* tumors, which also display variable degrees of differentiation and progression, ranging from early tumors to high-grade malignant, invasive carcinomas. In addition to the tumors detected in this study, a bilateral invasive carcinoma with clear-cell histology was detected within 12 months in an animal whose ovaries were treated bilaterally with ~5 μg of DMBA (not part of the three study arms). This advanced tumor displayed widespread dissemination to i.p. organs, production of ascites, and metastatic hemorrhagic foci in the lungs (data not shown).

Statistically, the appearance of lesions of any given severity did not depend significantly on the time of sacrifice after DMBA treatment; however, escalation of carcinogen dose combined with hormonal stimulation increased significantly the severity of the detected lesions. The cumulative incidence of preneoplastic lesions and tumors was also equivalently increased significantly at the higher DMBA dose in arms 2 and 3. Although the lesion incidence in arms 2 and 3 was similar, the lesions detected in arm 3 were more advanced than those in arm 2, including *bona fide* tumors that were not observed altogether in arm 2. This data demonstrates the strong contribution of gonadotropin hormones to the neoplastic progression of the ovarian lesions, perhaps due to increased ovarian surface epithelium cell proliferation and their effects on the underlying stroma. As demonstrated earlier, treatment of rats with pregnant mare's serum gonadotropin and/or human chorionic gonadotropin, in the presence or absence of surgical scarring to the ovary, leads to a 5 to 10-fold increase in the rate of ovarian surface epithelium cell proliferation (26).

The observed DMBA-induced reduction in ovarian volume, accompanied by decreased follicular growth and *corpora lutea* formation, is in good agreement with previously published data (28). The apparent differences in the observed low-dose response and persistence of ovarian hypoplasia in this study may be due to the slow-release form of DMBA applied directly to the ovary. Although not yet well understood in its full complexity, a suggested mechanism underlying the observed ovarian hypoplasia and cellular destruction is that DNA-adduct formation by DMBA metabolites leads to Tp53-mediated inhibition of DNA synthesis, cell growth arrest, and caspase-dependent or independent apoptosis (29–31). Hence, DMBA-induced mutation(s) that disrupt Tp53 function may allow evasion of affected ovarian surface epithelium cells and contribute to their malignant transformation.

Nonneoplastic and a small number of preneoplastic lesions, as well as a small granulosa cell tumor were also detected in control ovaries. To determine whether such lesions occur spontaneously in this rat strain, 20 nontreated animals were divided in two groups of 10 and maintained to the age of 8 and 14 months, respectively. Examination of their ovaries revealed no significant lesions, which strongly suggests that the lesions observed in the control ovaries may be a consequence of surgical scarring and chronic inflammation, and/or carcinogen carryover from the contralateral ovary. This data indicates that chronic inflammation, a known risk factor of ovarian cancer, may contribute to the DMBA-induced neoplastic process, either directly on epithelial cells through the action of secreted inflammatory cytokines and growth factors or indirectly through their effect on the adjacent stroma.

This study has additionally demonstrated that specific mutations in the *Tp53* and *Ki-Ras* genes, which are among the most frequent mutations found in human ovarian tumors, are also associated with ovarian cancer induced by DMBA. *TP53* mutations are found in 35 to 40% of human ovarian tumors (32–34). The identified rat *Tp53* mutations of codons 173 and 218 correspond to human codons 175 and 220, respectively, which are among the most frequent in human ovarian cancer (6.8% and 2.4, respectively).[3] Interestingly, both mutations lead to a characteristic accumulation of Tp53 protein. Activating mutations of *Ki-Ras*, including codon 61 detected in multiple DMBA-induced preneoplastic lesions and in one carcinoma, have been associated with ~20% of human ovarian tumors: of them, ~60% are found in mucinous and ~20% in serous carcinomas (35, 36). The relatively high frequency of *Ki-Ras* mutations in the preneoplastic lesions and, especially, in the ones with dysplasia provides a strong indication of their clonal (*i.e.,* neoplastic) nature. It additionally argues that *Ki-Ras* activation, either through mutation or by aberrant upstream signals, is very important during ovarian cancer development. Finally, a significant overexpression of the ER-α and PgR proteins was also demonstrated in the preneoplastic lesions and the serous low malignant potential tumor. However, the expression of the two receptors was markedly decreased or absent in the advanced carcinomas. The importance of this finding, in view of the existing controversy over the expression status of ER-α and PgR in human ovarian cancer (37, 38), mandates additional investigation. Furthermore, the Val^{660}Leu polymorphism that frequently occurs in exon 4 of PgRs has been suggested to have an association with human ovarian cancer characteristics and with overall ovarian cancer risk (39). Population-based studies, however, have demonstrated that no such association exists (40, 41). Lack of this PgR mutation in the examined ovarian lesions is additional evidence to the consistency of the DMBA rat ovarian cancer model with the human disease.

DMBA is a pluripotent carcinogen, which, through the formation of DNA adducts, induces initiating point mutations that alter the expression and/or activity of a number of oncogenes and tumor suppressor genes (42–45). Although DMBA itself is not a known environmental carcinogen associated with ovarian cancer, it shares similar mutagenic mechanisms with other polycyclic aromatic hydrocarbons whose abundance is relatively high in air pollutants and in tobacco smoke and which have been implicated in human cancer development (46, 47). Hence, the observed effect of DMBA in the ovary may be representative of the effect that such carcinogens have in the ovaries of affected women.

Here, we have demonstrated that direct application of a low dose of DMBA in the rat ovary, alone or combined with multiple cycles of gonadotropin administration, elicits a neoplastic process that affects mostly the ovarian surface epithelium and leads to the progressive development of putative epithelial cell preneoplasia, serous low malignant potential tumors, and invasive carcinomas. The similarity in histology and path of dissemination of the DMBA-induced rat ovarian carcinomas with those in the human, as well as the presence of gene mutations that are common in human ovarian cancer, demonstrate the validity of this animal model for additional delineation of the mechanisms underlying ovarian tumorigenesis. Finally, DMBA-induced ovarian oncogenesis in the rat could be used to preclinically test new agents for the prevention and/or therapy of the disease.

ACKNOWLEDGMENTS

We thank C. F. Renner for expert technical assistance and Drs. A. K. Godwin and X-X. Xu for their comments and suggestions.

REFERENCES

1. Gabra H, Smyth J. Biology of female cancers. New York: CRC Press; 1997.

[3] Internet address: http://www.iarc.fr/P53/index.html.

2. Ozols RF, Bookman MA, Connolly DC, et al. Focus on epithelial ovarian cancer. Cancer Cell 2004;5:19–24.
3. Dimascio J, Schilder RJ. Early stage management. In: Ozols RF, editor. Ovarian cancer: ACS atlas of clinical oncology. Hamilton, Ontario, Canada: BC Decker, Inc.; 2002. p. 147–58.
4. Ozols RF. Primary Chemotherapy Regimens. In: Ozols RF, editor. Ovarian cancer: ACS atlas of clinical oncology. Hamilton, Ontario, Canada: BC Decker, Inc.; 2002. p. 119–31.
5. Salazar H, Godwin AK, Daly MB, et al. Microscopic benign and invasive malignant neoplasms and a cancer-prone phenotype in prophylactic oophorectomies [see comments]. J Natl Cancer Inst (Bethesda) 1996;88:1810–20.
6. Stratton JF, Buckley CH, Lowe D, Ponder BA. Comparison of prophylactic oophorectomy specimens from carriers and noncarriers of a BRCA1 or BRCA2 gene mutation. J Natl Cancer Inst (Bethesda) 1999;91:626–8.
7. Resta L, Russo S, Colucci GA, Prat J. Morphologic precursors of ovarian epithelial tumors. Obstet Gynecol 1993;82:181–6.
8. Scully RE. Pathology of ovarian cancer precursors. J Cell Biochem Suppl 1995;23: 208–18.
9. Auersperg N, Wong AS, Choi KC, et al. Ovarian surface epithelium: biology, endocrinology, and pathology. Endocr Rev 2001;22:255–88.
10. Mossman HW, Duke KL. Comparative morphology of the mammalian ovary. Madison, WI: University of Wisconsin Press; 1997.
11. Russell P. The pathological assessment of ovarian neoplasms. III: The malignant "epithelial" tumours. Pathology 1979;11:493–532.
12. Wynder EL, Dodo H, Barber HR. Epidemiology of cancer of the ovary. Cancer (Phila.) 1969;23:352–70.
13. Fathalla MF. Incessant ovulation: a factor in ovarian neoplasia? Lancet 1971;2:163.
14. Fathalla MF. Factors in the causation and incidence of ovarian cancer. Obstet Gynecol Surv 1972;27:751–68.
15. Hamilton TC, Xu X-X, Patriotis C, Salazar H. Biology. In: Ozols RF, editor. Ovarian cancer: ACS atlas of clinical oncology. Hamilton, Ontario, Canada: BC Decker, Inc.; 2002. p. 27–38.
16. Ness RB, Cottreau C. Possible role of ovarian epithelial inflammation in ovarian cancer. J Natl Cancer Inst (Bethesda) 1999;91:1459–67.
17. Stakleff KD, Von Gruenigen VE. Rodent models for ovarian cancer research. Int J Gynecol Cancer 2003;13:405–12.
18. Hamilton TC, Connolly DC, Nikitin AY, et al. Translational research in ovarian cancer: a must. Int J Gynecol Cancer 2003;13(Suppl 2):220–30.
19. Flesken-Nikitin A, Choi KC, Eng JP, et al. Induction of carcinogenesis by concurrent inactivation of p53 and Rb1 in the mouse ovarian surface epithelium. Cancer Res 2003;63:3459–63.
20. Orsulic S, Li Y, Soslow RA, et al. Induction of ovarian cancer by defined multiple genetic changes in a mouse model system. Cancer Cell 2002;1:53–62.
21. Connolly DC, Bao R, Nikitin AY, et al. Female mice chimeric for expression of the simian virus 40 TAg under control of the MISIIR promoter develop epithelial ovarian cancer. Cancer Res 2003;63:1389–97.
22. Sekiya S, Endoh N, Kikuchi Y, et al. In vivo and in vitro studies of experimental ovarian adenocarcinoma in rats. Cancer Res 1979;39:1108–12.
23. Tunca JC, Erturk E, Bryan GT. Chemical induction of ovarian tumors in rats. Gynecol Oncol 1985;21:54–64.
24. Kato T, Yakushiji M, Tunawaki A, Ide K. A study of experimental ovarian tumors in rats by chemical carcinogen, 20-methylcholanthrene. Kurume Med J 1973;20:159–67.
25. Nishida T, Sugiyama T, Kataoka A, et al. Histologic characterization of rat ovarian carcinoma induced by intraovarian insertion of a 7,12-dimethylbenz[alpha]anthracene-coated suture: common epithelial tumors of the ovary in rats? Cancer 1998;83: 965–70.
26. Stewart SL, Querec TD, Gruver BN, et al. Gonadotropin and steroid hormones stimulate proliferation of the rat ovarian surface epithelium. J Cell Physiol 2004;198: 119–24.
27. Bennett JM, Catovsky D, Daniel MT, et al. A variant form of hypergranular promyelocytic leukaemia (M3). Br J Haematol 1980;44:169–70.
28. Weitzman GA, Miller MM, London SN, Mattison DR. Morphometric assessment of the murine ovarian toxicity of 7,12-dimethylbenz(alpha)anthracene. Reprod Toxicol 1992;6:137–41.
29. Page TJ, O'Brien S, Jefcoate CR, Czuprynski CJ. 7,12-Dimethylbenz[alpha]anthracene induces apoptosis in murine pre-B cells through a caspase-8–dependent pathway. Mol Pharmacol 2002;62:313–9.
30. Page TJ, O'Brien S, Holston K, et al. 7,12-Dimethylbenz[alpha]anthracene-induced bone marrow toxicity is p53 dependent. Toxicol Sci 2003;74:85–92.
31. Tsuta K, Shikata N, Kominami S, Tsubura A. Mechanisms of adrenal damage induced by 7,12-dimethylbenz(alpha)anthracene in female Sprague-Dawley rats. Exp Mol Pathol 2001;70:162–72.
32. Kohler MF, Kerns BJ, Humphrey PA, et al. Mutation and overexpression of p53 in early-stage epithelial ovarian cancer. Obstet Gynecol 1993;81:643–50.
33. Kupryjanczyk J, Thor AD, Beauchamp R, et al. p53 gene mutations and protein accumulation in human ovarian cancer. Proc Natl Acad Sci USA 1993;90:4961–5.
34. Wang Y, Helland A, Holm R, et al. TP53 mutations in early-stage ovarian carcinoma, relation to long-term survival. Br J Cancer 2004;90:678–85.
35. Enomoto T, Weghorst CM, Inoue M, et al. K-ras activation occurs frequently in mucinous adenocarcinomas and rarely in other common epithelial tumors of the human ovary. Am J Pathol 1991;139:777–85.
36. Chien CH, Chow SN. Point mutation of the ras oncogene in human ovarian cancer. DNA Cell Biol 1993;12:623–27.
37. Lau KM, Mok SC, Ho SM. Expression of human estrogen receptor-alpha and -beta, progesterone receptor, and androgen receptor mRNA in normal and malignant ovarian epithelial cells. Proc Natl Acad Sci USA 1999;96:5722–7.
38. Li AJ, Baldwin RL, Karlan BY. Estrogen and progesterone receptor subtype expression in normal and malignant ovarian epithelial cell cultures. Am J Obstet Gynecol 2003;189:22–7.
39. Lancaster JM, Berchuck A, Carney ME, et al. Progesterone receptor gene polymorphism and risk for breast and ovarian cancer. Br J Cancer 1998;78:277.
40. Tong D, Fabjani G, Heinze G, et al. Analysis of the human progesterone receptor gene polymorphism progins in Austrian ovarian carcinoma patients. Int J Cancer 2001;95: 394–7.
41. Spurdle AB, Webb PM, Purdie DM, et al. No significant association between progesterone receptor exon 4 Val660Leu G/T polymorphism and risk of ovarian cancer. Carcinogenesis (Lond.) 2001;22:717–21.
42. Osaka M, Matsuo S, Koh T, Sugiyama T. Loss of heterozygosity at the N-ras locus in 7,12-dimethylbenz[alpha] anthracene-induced rat leukemia. Mol Carcinog 1997; 18:206–12.
43. Osaka M, Koh T, Matsuo S, Sugiyama T. The specific N-ras mutation in rat 7,12-dimethylbenz[alpha]anthracene (DMBA)-induced leukemia. Leukemia (Baltimore) 1997;11(Suppl 3):393–5.
44. Kito K, Kihana T, Sugita A, et al. Incidence of p53 and Ha-ras gene mutations in chemically induced rat mammary carcinomas. Mol Carcinog 1996;17:78–83.
45. Ember I, Kiss I, Pusztai Z. Effect of 7,12-dimethylbenz(alpha)anthracene on onco/suppressor gene action in vivo: a short-term experiment. Anticancer Res 1998;18: 445–7.
46. Arif JM, Smith WA, Gupta RC. Tissue distribution of DNA adducts in rats treated by intramammillary injection with dibenzo[a,l]pyrene, 7,12-dimethylbenz[a]anthracene and benzo[a]pyrene. Mutat Res 1997;378:31–9.
47. Vainio H, Matos E, Kogevinas M. Identification of occupational carcinogens. IARC Scientific Publ. No. . Lyon, France: IARC; 1994;129:41–59.

BIOINFORMATICS

Vol. 20 no. 11 2004, pages 1772–1784
doi:10.1093/bioinformatics/bth170

Normalization of single-channel DNA array data by principal component analysis

Radka Stoyanova[1], Troy D. Querec[2,†], Truman R. Brown[3] and Christos Patriotis[2,*]

[1]Division of Population Science and [2]Department of Medical Oncology,
Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111-2497 and
[3]Hatch Center for MR Research, Columbia University, 710 W. 168th St., New York, NY 10032, USA

Received on July 9, 2003; revised on January 8, 2004; accepted on February 8, 2004
Advance Access publication March 22, 2004

ABSTRACT

Motivation: Detailed comparison and analysis of the output of DNA gene expression arrays from multiple samples require global normalization of the measured individual gene intensities from the different hybridizations. This is needed for accounting for variations in array preparation and sample hybridization conditions.

Results: Here, we present a simple, robust and accurate procedure for the global normalization of datasets generated with single-channel DNA arrays based on principal component analysis. The procedure makes minimal assumptions about the data and performs well in cases where other standard procedures produced biased estimates. It is also insensitive to data transformation, filtering (thresholding) and pre-screening.

Contact: Christos.Patriotis@fccc.edu

INTRODUCTION

The development of high-density DNA arrays (oligonucleotide and cDNA) has revolutionized our ability to characterize biological processes and samples genetically by monitoring the relative expression of thousands of genes simultaneously (Bowtell, 1999; Debouck and Goodfellow, 1999; Duggan *et al*., 1999; Lander, 1999). To meet the challenges for interpretation of this complex data, sophisticated software packages have become available for analysis of the gene expression profiles, such as ScanAnalyze (Eisen and Brown, 1999), ArrayExplorer (Patriotis *et al*., 2001) and ImaGene (Biodiscovery, Inc.). An important, but still unresolved, issue is associated with the normalization of the relative expression of genes across a series of microarray experiments. In order to compare the results from multiple samples, which is the ultimate goal of these studies, it is obligatory that the individual array datasets be normalized to correct for the inherent experimental differences. The critical element in this process is the discrimination of the interesting, biological variation from the obscuring variation, which is related to the experimental conditions (Hartemink *et al*., 2001). This is why the initial attempts towards normalization of array datasets relied on the concept that a group of genes could be identified *a priori* and serve as ‘housekeeping’ genes, assuming that their expression will reflect directly the obscuring experimental variation. As discussed in detail below, if such a subset of genes could be identified reliably, then well-defined normalization factors could be estimated to within the accuracy inherent in the measurements. Unfortunately, as shown by others (Butte *et al*., 2001; Selvey *et al*., 2001) and by us in this report, this simple concept works only in very limited cases. (Here and in the rest of the paper, we will refer to the *a priori* specified housekeeping genes as ‘designated’ in order to distinguish them from those determined to be the ‘true’ housekeeping genes. The latter represent the subset of genes whose expression is invariant to the particular biological and/or experimental variables in the multiple microarray experiments being compared.)

The realization that in most of the cases the ‘designated’ housekeeping genes cannot be used for reliable normalization has spurred the development of alternative approaches for normalization. The majority of these approaches determine normalization factors on the basis of averages over the behavior of the entire set of genes measured (Schuchhardt *et al*., 2000). Typically, these methods utilize the mean or median of the array intensities (Quackenbush, 2001) and linear (Golub *et al*., 1999) or orthogonal regression (Sapir and Churchill, 2000). A variety of non-linear techniques were also proposed (Schadt *et al*., 2000, 2001; Li and Wong, 2001; Bolstad *et al*., 2003).

There is also a series of methods that identify a subset of genes in the data that can be assumed as housekeeping (Zien *et al*., 2001; Kepler *et al*., 2002). All these approaches perform

*To whom correspondence should be addressed.

†Present address: Emory University, GDBBS, 1462 Clifton Road, Dental Bldg, Suite 314, Atlanta, GA 30322, USA.

satisfactorily when the following two assumptions about the data are met:

(1) the majority of the genes (in the fitting segment for the non-linear approaches, or overall) are not affected by the experimental variables, i.e. they can all be regarded as housekeeping genes; and

(2) the subset of differentially expressed genes are 'activated' symmetrically, i.e. the overall intensity change of up- and down-regulated genes is similar.

Here we present a novel normalization approach that performs satisfactorily even when the conditions above are not met, which is the most commonly observed scenario. In contrast to the methods requiring the selection of a baseline array, this method analyses the entire dataset simultaneously, and, as such, it is considered a complete data method (Bolstad *et al*., 2003). The goal of the technique is to determine in a multi-array experiment if there is a subset of genes whose expression may be considered unaffected by the 'interesting' (biological) sources of variation and if there are such, to identify this set of specific, 'data-driven' housekeeping genes and use them for normalization. Briefly, if the results from each array measurement are represented in a multi-dimensional vector space where each axis is a different sample, then the entire experiment can be represented as a series of points corresponding to the strength of each gene's expression in each sample measured. If a set of genes with an unchanged relative expression is present, their intensity levels will represent points along a straight line through the origin. We present a principal component analysis (PCA)-based method for identifying such a line, if one exists. The factors determined from the expression of these genes can be used to normalize the gene expression in the individual array datasets.

MATERIALS AND METHODS

Theory

Consider a gene expression dataset consisting of m arrays with n genes each. Let $\mathbf{D}$ be the data matrix containing in its rows the measured expression levels, and let g_{ij} be the measured expression level of the i-th gene in the j-th array ($i = 1, \dots, n, j = 1, \dots, m$). We seek to identify a subset, $\mathbf{S}$, of s genes ($s \leq n$) whose expression remains constant over the experimental conditions of the study. Mathematically, for the genes in S the following equations hold:

$$q_j g_{ij} = c_i \quad \text{or} \quad g_{ij} = c_i / q_j,$$

where q_j is the j-th normalization constant and c_i is the true concentration of the i-th gene, which is constant across the samples. If we plot the points g_{ij} in an m-dimensional space, we can see that they lie along a line through the origin, which has projections along the axes of $\{1/q_j\}$. If we can find such a line, we will have identified our desired relative normalization constants (relative since unless at least one of the c_is is known, it is impossible to normalize the data absolutely).

We now turn to the problem of identifying the genes in $\mathbf{S}$. The obvious method is to calculate the densities in the cloud of n data points in the m-dimensional data space, which represent the directions of n gene levels in the m observations. In reality, this is difficult because there are approximately N^{m-1} directions for examining if each orientation is divided into N segments. In order to reduce the dimensions of the space that needs to be examined, we use PCA to identify the directions along which the principal variations of the genetic expressions lie in the original m-dimensional space. We project the data points onto the first two of these directions and examine their angular distribution to determine if a line through the origin is present. Note that the original line in the full space need not lie in this plane as its projection into the plane will also be a line through the origin.

PCA is used commonly for reducing the dimensionality of complex data (Anderson, 1971) and has been used previously in the analysis of microarray data from time-course experiments (Alter *et al*., 2000, 2003), for normalization of gene expression ratios obtained from two different microchips of two-channel arrays (Nielsen *et al*., 2002) and for partitioning large-sample microarray-based gene expression profiles (Peterson, 2003). It is also an inseparable part for exploration of large genomic datasets (Misra *et al*., 2002). Previously, we have applied the PCA technique for removing 'unwanted' variation in multi-spectral datasets (Stoyanova and Brown, 2002).

Briefly, PCA identifies the directions of the largest variations in the data via the principal components (PCs), and represents the data in a coordinate system defined by the PCs ($\vec{P}_1, \vec{P}_2, \dots$), as follows:

$$\mathbf{D} = R_1\vec{P}_1 + R_2\vec{P}_2 + R_3\vec{P}_3 + \cdots + R_m\vec{P}_m, \qquad (1)$$

where $\vec{P}_j$ $(1 \times m)$ and R_j $(n \times 1)$ are row and column matrices; R_j contain the projections of the data along the PCs ($j = 1, \dots, m$), generally called scores. Below, some of the relevant properties of the PCs are listed.

(1) $\vec{P}_j$ are eigenvectors of the data-covariance matrix (calculated around the origin, rather than around the mean) and are orthonormal, i.e.

$$\vec{P}_i \cdot \vec{P}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$$

(2) The PCs are ordered by the decreasing amount of variation in the data they explain. Let $\Lambda_1, \Lambda_2, \dots, \Lambda_m$ be the eigenvalues of the covariance matrix ($\Lambda_1 > \Lambda_2 > \cdots > \Lambda_m$). Each PC explains a portion of the total variance of $\mathbf{D}$, proportional to its corresponding eigenvalue.

(3) The magnitude of R_j is proportional to its corresponding eigenvalue, Λ_j.

Cvetkovic, Dusica

(4) **D** can be represented sufficiently with fewer than m PCs [Equation (1)]. PCA provides a representation of the data in a lower-dimensional space of significant variables.

(5) The PCs are a linear combination of the original data. The coefficients of this linear combination (R_i) are typically referred to as loadings and represent the projections of the PCs along the axes of the original m-dimensional space.

(6) The PCs minimize the squared distances of the variables (gene-expression levels) and themselves.

From the last three properties, it follows that the loadings of the first PC may serve as normalization coefficients of the arrays. In many cases, when the assumptions (1) and (2) (see Introduction) are met, as discussed in detail below, PCA can provide directly the normalization coefficients sought. In other cases, we can use the first two PCs to detect linear behavior in a subset of genes **S** ($s \leq n$) that are the 'true' housekeeping genes. PCA applied only to the genes in **S** will identify the appropriate normalization line in the entire m-dimensional data space. Its projections can then be used as normalization factors.

The procedure [dubbed PCA(line)] tests automatically for the existence of and detects the group of genes, which are distributed 'tightly' along a line in the plane defined by the first two PCs. We chose this plane because by definition it contains the largest variations in the expression levels. Although the actual straight line of the desired normalization may not lie completely in this plane, its projection in the plane is also a straight line and will serve to identify the desired set of genes. To identify such a line, we divide the part of the plane that contains all the points into small angular segments and determine the number of data points (genes) in each segment. The segment(s) containing the data-driven housekeeping genes will contain a disproportionally large density of points. This procedure is described below and given in detail in Appendix 1.

Initially, we assume **S** is an empty set (**S** $\equiv \emptyset$). In the plane defined by $\vec{P}_1$ and $\vec{P}_2$, we partition the angle through the origin defined by the genes with maximal and minimal components on $\vec{P}_2$ in p equal angular segments. Let s_k ($k = 1, \ldots, p$) be the subset of genes in **D**, that belong to the k-th segment ($s_1 \cup s_2 \cup \cdots \cup s_p =$ **D**). We recommend that p be set initially to contain on average at least 10 genes per segment. Let θ_k be the angular densities defined as the number of genes in each segment, s_k, and $M(\theta_k)$ and $V(\theta_k)$ be, respectively, the sample mean and variance of θ_k. Then, the density of the k-th segment is considered to be significant if

$$\theta_k > M(\theta_k) + \mu\sqrt{V(\theta_k)}, \qquad (2)$$

where μ is a parameter indicating the number of standard deviations above the mean that is required for significance. If a normal distribution of θ_k is assumed, then $\mu = 1.96$ will correspond to a one-sided test with a type-I error of 2.5%. However, in most cases, due to different procedures for microarray image quantification as well as the specific prefiltering of the data, the distribution of θ_k is unknown. In cases where a normal distribution of θ_k cannot be assumed, it is recommended that their histogram be examined and μ be set appropriately. For added stringency of the test, the genes in segment s_k are assumed to be housekeeping genes only if θ_{k+1} of the neighbouring segment s_{k+1} is also tested significant. Then the genes in the two segments are merged in **S**, i.e. **S** $\equiv s_k \cup s_{k+1}$. If the angular density of the genes of further contiguous segments is detected to be significant, then these genes are added to **S**. After all segments are tested, PCA is applied to **S** and the reciprocal values of the loadings of the resultant first PC are used as normalization coefficients.

If the procedure failed to identify at least two significant contiguous segments, then either all the genes in the data can be assumed to be housekeeping (**S** $\equiv$ **D**), or, in the extreme situation, the housekeeping genes are either too few to be detected or not existent (**S** $\equiv \emptyset$). In the first case, the loadings of the first PC from the initial PCA of **D** are the true normalization coefficients and can be used for direct normalization. There is not very much to be done in the second case—the PCA-derived normalization would be as erroneous as the ones produced by any other linear technique. Let λ_1 be the fraction (in per cent) of the first eigenvalue, Λ_1, from the total variance in the data. In this case, a low λ_1 (in our experience <60%) will be indicative of a lack of normalizing genes.

Biological samples (datasets)

Human ovarian surface epithelial cell lines Microarray datasets obtained from experiments with RNA of human ovarian surface epithelial (HOSE) cells were analyzed using Atlas 1.2 Human arrays (ClonTech). The details of array preparation and data extraction are described elsewhere (Patriotis *et al.*, 2001). Briefly, the HOSE cells were derived from a short-term primary cell culture obtained from one of the ovaries of an individual predisposed to ovarian cancer. The short-term HOSE cell culture was transduced with a Cytomegalovirus-based vector expressing the Simian Virus-40 large T-antigen. As a result, the *in vitro* lifespan of the cells, while still 'mortal' (118M), was considerably extended, leading to the spontaneous outgrowth of an 'immortal'/non-transformed cell line (118Im). Following multiple passages in culture, the 118Im cell line gave rise spontaneously to cells that acquired anchorage-independent growth characteristics and, ultimately, the potential to grow tumours *in vivo* when inoculated in nude mice (118NuTu) (Frolov, A. *et al.*, unpublished data). In the first experiment, the cDNA probes were derived from total RNA purified from 118M, 118Im and 118NuTu. In the second experiment, microarray data were obtained from 118NuTu cells treated for different lengths of time (0, 24, 48 and 72 h) with the synthetic retinoic acid derivative Fenretinide (4-HPR) (Moon *et al.*, 1979).

Lymphoma data (LD)

The dataset was constructed from the supplementary datasets of Golub *et al.* (1999). The microarray measurements were performed with RNA of samples obtained from bone marrow and peripheral blood from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML) at the time of diagnosis using high-density oligonucleotide Affymetrix arrays. In the paper referred to, the data were normalized by pair-wise linear regression (LR) between the first sample (baseline) and the rest of the samples in the dataset. Only genes with satisfactory quality (marked with 'P' in the datasets provided) in each pair were considered for the regression. The normalized datasets, as well as the normalization factors, are supplied at http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi. The data used here were non-processed and 'non-normalized', and the combined datasets resulted in a data matrix containing 72 arrays and 7129 genes.

Simulated data

The values in the simulated datasets were chosen to be realistically probable, based on our experience with data obtained with the Atlas 1.2 CLONTECH arrays (Patriotis *et al.*, 2001). The number of genes was set to 500, in agreement with our observation that between 30 and 50% of the genes are expressed in any of the samples investigated in our lab. In the first array, the expression levels, g_{i1} [in arbitrary units (a.u.)], were simulated using the relation $g_{i1} = 2^u$, where u is uniformly distributed between 1 and 16.

In all simulated datasets of pairs of arrays a multiplication factor of 1.2 was applied to the second array, equivalent to $q_1 = 1$ and $q_2 = 1.2$. Gene intensities were assumed to be background-corrected, and (unless noted otherwise) signals with intensities less than 200 were zeroed (thresholded).

'Noise' data

The sources of noise in microarray datasets are multiple and complex, and they contribute simultaneously with variable amounts to the total variance in the data. Generally, the total noise contribution to the measured signal represents a variable mixture of the contribution of two components: one is independent of gene intensity and affects the expression of all genes equally, and the other is gene-dependent and increases with the magnitude of the gene expression. To investigate the contribution of noise to the process of normalization, we simulated two pairs of replicate arrays, as described above. Random noise was added to each array. In the first set, the noise was gene independent (N_1)—uniformly distributed random noise between −2500 and 2500—and in the second set, a gene-dependent (N_2), uniformly distributed noise whose magnitude was ±10% of the gene intensities. Formally,

$$\begin{aligned} N_1 &= -2500 + 5000\,u \\ N_2 &= \frac{g_{i1,2}}{10}(2u-1) \end{aligned} \qquad u = U(0,1). \tag{3}$$

'Signal' dataset 1

'Signal' dataset 1 (SD1) contained two pairs of simulated arrays. The first pair satisfied conditions (1) and (2) (see Introduction) by choosing a substantial number of the genes to be housekeeping (250) and the number and magnitude of change of up- and down-regulated genes to be equal. The second pair was constructed to illustrate a scenario where these assumptions are not met: the housekeeping genes (150) were not a majority, and more genes were 'up-regulated' (200) than 'down-regulated' (150) (the details about the simulated up- and down-regulation are given in Appendix 2). Two independent sets of random noise were added to each array, generated as the sum of half of both gene-dependent and -independent noise [Equation (3)], i.e. $\frac{1}{2}(N_1 + N_2)$.

'Signal' dataset 2

'Signal' dataset 2 (SD2) contained eight arrays with 500 genes each. The first array in SD2 was generated randomly, as described above. The gene expression levels of the remaining seven arrays were generated with the idea of recreating a scenario where progressive changes occur in the studied samples (e.g. time-response to treatment or undergoing a process of immortalization and malignant transformation). The details of simulation parameters for up- and down-regulation are given in Appendix 3. The arrays were multiplied with coefficients generated at random between 0.3 and 3. Finally, random noise, generated as described for SD1, was added to each array.

RESULTS

Housekeeping genes in HOSE cells

Figure 1(a) depicts the correlation plot of the 'designated' housekeeping genes in the first experiment with HOSE cells: 118M on the x-axis, and on the y-axis 118Im (black series) and 118NuTu (gray series). The expression of these genes is well correlated ($R^2 = 0.96$), and, in this case, they can be used for normalization of the data. Figure 1(b) depicts the correlation plot of the expression of the same set of housekeeping genes in the 118NuTu, untreated (0 h, x-axis) and treated with 4-HPR for 24, 48 and 72 h (y-axis; black circles, gray triangles and shaded squares, respectively). In this case, the correlation between the expression of the 'designated' housekeeping genes is quite poor ($R^2 = 0.43$, 0.81 and 0.85, respectively). From these data, it is clear that the expression profiles of the 'designated' housekeeping genes are changed non-uniformly in the cells in response to the drug treatment.

'Noise' data

Figure 2(a) and (b) (left panels) depict the correlation between the data in the two pairs of simulated arrays in this dataset together with the linear trendline through the origin. Note that the regression coefficient in both cases is very close to the true value of the multiplication factor 1.2. The fit is slightly tighter

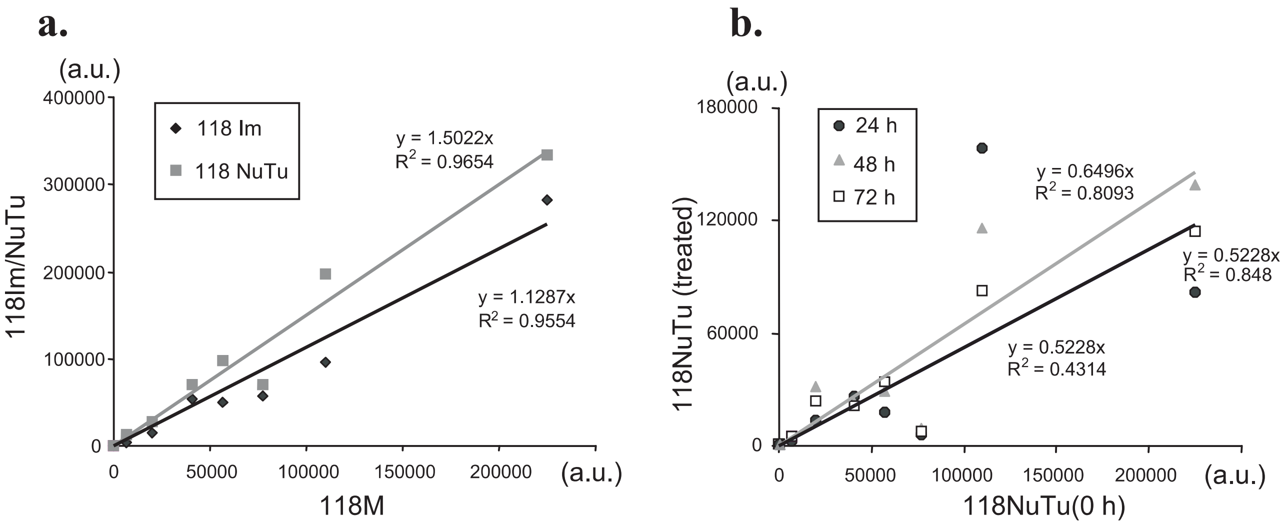

Fig. 1. Correlation plots of the intensities of the 'designated' housekeeping genes in two microarray experiments. (**a**) HOSE cell lines at different stages of malignancy, on the x-axis 118M, and on the y-axis, 118Im (black) and 118NuTu (gray). Regression lines are indicated in black and gray, respectively; (**b**) 118NuTu cell line following treatment with Fenretinide, on the x-axis at 0 h and on the y-axis after 24 (black circles), 48 (gray triangles) and 72 h (squares) of treatment. Regression lines are indicated in black solid, black dashed and gray, respectively (note that the black solid and black dashed regression lines are overlapping).

for the second dataset ($R^2 = 0.986$ versus $R^2 = 0.992$), which reflects the smaller contribution of the noise in the overall gene intensities. Figure 2(c) (left panel) depicts the correlation between two replicate array datasets obtained from 118M. The genes depicted by gray squares represent the 'designated' housekeeping genes. On the right panels in Figure 2 the correlation of the logarithmic transforms of the data from the left panels are presented (due to the restriction of the logarithmic function to only positive numbers, for this comparison, only genes that are expressed simultaneously in the two arrays are used). Comparison of the graphs of simulated [Fig. 2(a) and (b)] and real [Fig. 2(c)] noise indicates the similarity in the overall distributions, although the real data have a greater variance.

'Signal' dataset SD1

The graphs of the two pairs of arrays in this dataset, together with the regression line through the origin, are presented in Figure 3. The housekeeping genes are marked in green. In the case of the first pair [Fig. 3(a)], it is clear that the regression line is along the line of normalization and, therefore, all the above reference normalization methods will perform well. Obviously, this is not the case with the second dataset [Fig. 3(b)], and we applied the PCA (line) procedure for determining the subset of housekeeping genes.

After thresholding, 296 genes were found with non-zero intensities simultaneously in both arrays (132 up-regulated, 88 down-regulated and 76 housekeeping). PCA was applied to this set ($\lambda_1 = 96\%$). The representation of the data along the first two PCs is shown in Figure 4(a) [note that the first PC, $\vec{P}_1$, is along the regression line of this rotated version of Fig. 3(b)]. The procedure for automatic detection of the housekeeping genes is schematically illustrated in Figure 4(b). The angle encompassing all data points (between 1.069 and 2.438 radians) was divided into 50 segments. The histogram of the angular densities θ_k ($k = 1, 2, \ldots, 50$) is presented in Figure 4(c) [$M(\theta_k) = 5.92$ and $\sqrt{V(\theta_k)} = 5.18$]. For $\mu = 1.96$, three contiguous segments, starting at $p = 22$, contained points with a significantly higher density [Equation (2)]. A total of 63 points (subset **S**) from these segments were extracted. These genes (orange points), together with the original set of housekeeping genes (in green), are presented in Figure 4(d). The collinearity between the identified genes and the housekeeping genes is apparent. Thirty-two of the genes in **S** belong to the original set of 76 housekeeping genes in the analyzed data, indicating that the procedure recovered successfully a substantial fraction of them (32/76, or >40%). Moreover, the procedure detected an additional 31 genes whose expression changes in accordance with a housekeeping gene behavior. PCA was applied to the data in **S** ($\lambda_1 = 99\%$), and the first PC loading factors were $q_1 = 0.635$ and $q_2 = 0.773$, corresponding to a relative normalization factor of 1.217.

Simulated dataset SD2

PCA was applied to 205 genes with non-zero intensities in all eight arrays (88 up-regulated, 52 down-regulated and 64 housekeeping) ($\lambda_1 = 96\%$). The points in the $\vec{P}_1$ and $\vec{P}_2$ plane were within 1.079 and 1.938 radians. As in the case of SD1, the densities of points in 50 segments were calculated ($M(\theta_k) = 4.08$ and $\sqrt{V(\theta_k)} = 5.21$). For $\mu = 1.96$, three

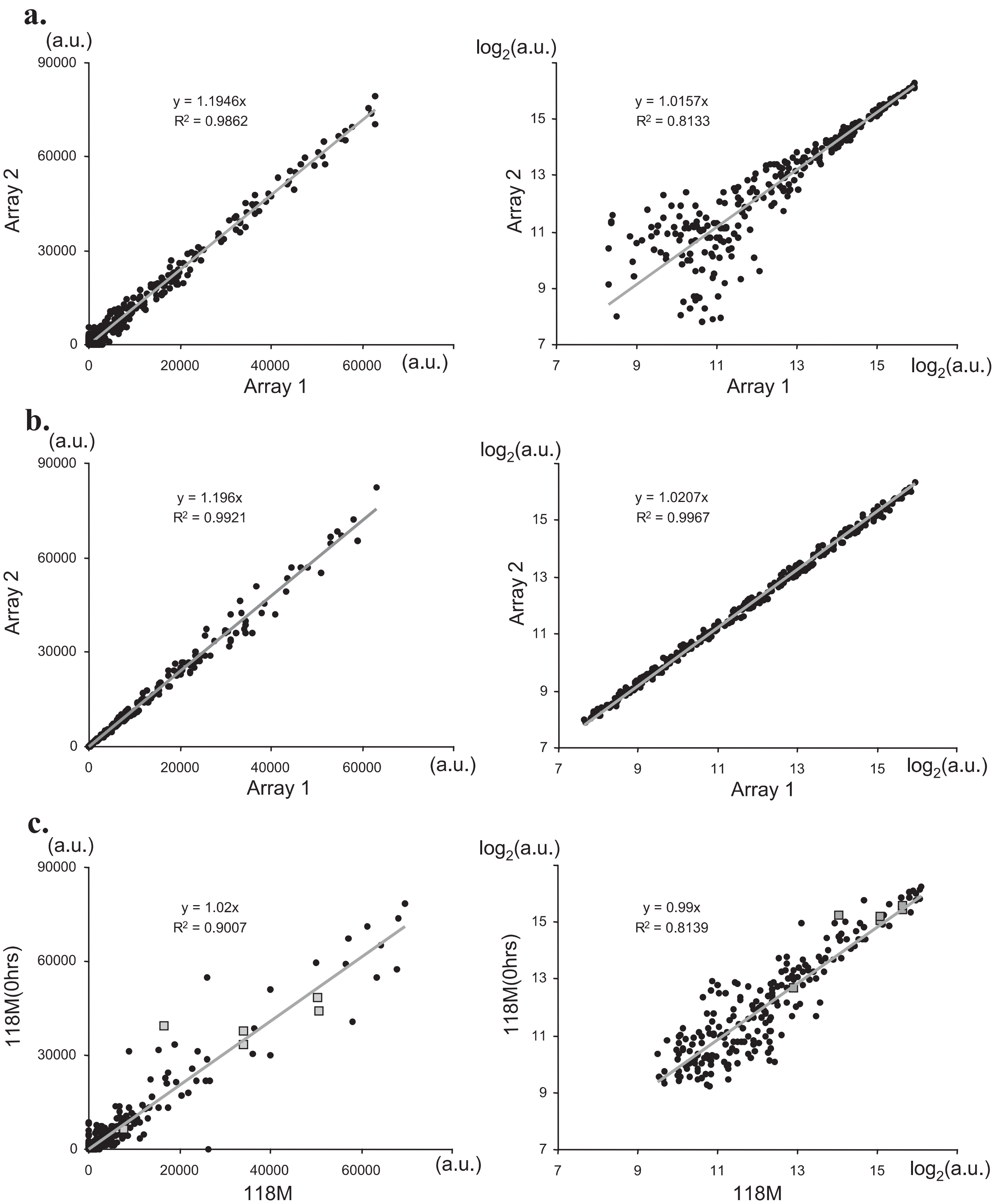

Fig. 2. Correlation plots of gene intensities in replicate arrays, displayed on untransformed (left panels) and logarithmic scales (right panels) with indicated LR line (gray): (**a**) simulated data, containing gene-independent noise; (**b**) simulated data, containing gene intensity-dependent noise; (**c**) two replicate arrays of 118M cell line. The genes shown in gray squares represent the designated housekeeping genes included in the arrays by the manufacturer.

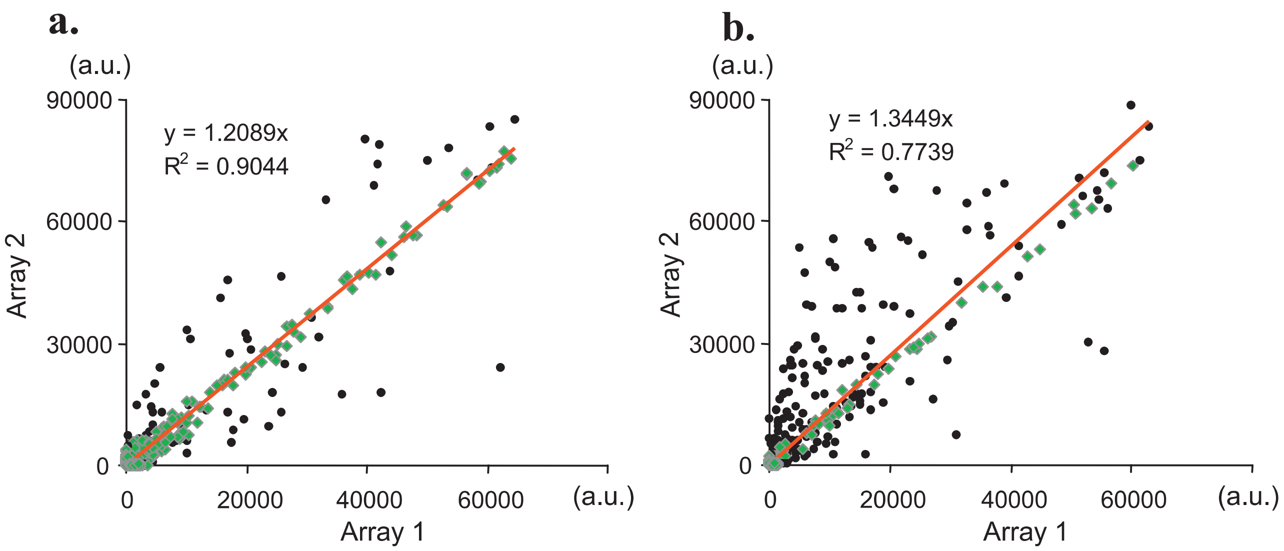

Fig. 3. Correlation plots of gene intensities of two simulated array datasets (SD1) with indicated housekeeping genes (green squares) and indicated LR line (orange): (**a**) 'symmetric' case, where the majority of the genes are housekeeping and the number and magnitude of up- and down-regulated genes is similar; (**b**) the housekeeping genes are of a relatively smaller number, and the up-regulated genes dominate the distribution.

contiguous segments containing a total of 64 points (subset **S**) contained a significant number of points. The majority of the points in **S** belonged to the original set of housekeeping genes analyzed (44, or 69%), and the remaining 20 were split between the 12 up-regulated and eight down-regulated genes. PCA was applied to the data in **S** ($\lambda_1 = 99\%$), and the normalization coefficients $q_j (j = 1, \ldots, 8)$ were calculated as the loadings of the first PC.

We compared the accuracy of the PCA(line)-estimated normalization factors with the ones estimated by LR and mean (MEAN). We scaled all normalization factors so that their sum was equal to 1, and the correlation between the true values (x-axis) and the estimated values (y-axis) are presented in Figure 5(a). Although the overall correlation between the true and estimated normalization factors is quite good [$R^2 = 0.9964$, 0.9862 and 0.9726 for PCA(line), LR and MEAN estimates, respectively], it is clear that PCA(line) provides the best estimates. We also calculated the error for each individual array, defined as the percentage difference of the estimated from the true normalization factor, and the minimum, maximum and average error values are presented in Figure 5(b). This analysis indicated that the error of the PCA(line)-derived estimates is on average lower by a factor of 2 and 3 as compared with the ones derived by LR and MEAN, respectively.

We further investigated the effect of data thresholding on the PCA(line) procedure. We re-analyzed SD2 by applying PCA to all 500 genes in the dataset. Since some of the scores along $\vec{P}_2$ were negative, the data points spanned the entire plane (between 0.03 and 6.27 radians). In this case, we set $p = 200$ and $\mu = 4$. Two consecutive segments [Fig. 5(c)], containing a total of 77 genes, were determined to have significant angular densities. The overwhelming majority of genes (55) in this set belonged to the original set of housekeeping genes. The housekeeping gene sets derived by PCA (line) on thresholded and unfiltered data were strongly overlapping—all but four were identical to the 64 housekeeping genes determined with the thresholded data. Finally, the PCA-determined normalization factors in this case were virtually identical to the ones determined with the thresholded data.

Lymphoma Data

PCA was applied to all 7129 genes in the dataset ($\lambda_1 = 88.31\%$). All loadings of $\vec{P}_1$ were scaled by the first one, resulting in a normalization factor of 1 for the first array. Figure 6(a) depicts the comparison between LR- and PCA-derived (yellow circles) values. The high correlation ($R^2 = 0.99$) between the two series is apparent. Further, we applied the PCA(line) procedure. Three contiguous segments (from a total of 200), containing 1095 genes, were above the threshold [$M(\theta_k) = 35.64$, $\sqrt{V(\theta_k)} = 72.21$, $\mu = 4$]. PCA was applied to the intensities of the genes in **S** ($\lambda_1 = 93.85\%$) and the loadings of $\vec{P}_1$ rescaled appropriately and compared with the LR results [Fig. 6(a), black circles]. While showing an overall good agreement with the LR-derived results ($R^2 = 0.92$), they also indicate, in some individual cases, substantial differences with the PCA(line)-estimated values. The average absolute value of the relative difference between LR- and PCA-derived factors was 7.52%, with a range of 0.07–30.84% in the case of array #65 [Fig. 6(a), marked with an arrow]. We then examined the correlation of the intensities of the genes marked with 'P' (those of satisfactory quality) in arrays # 1

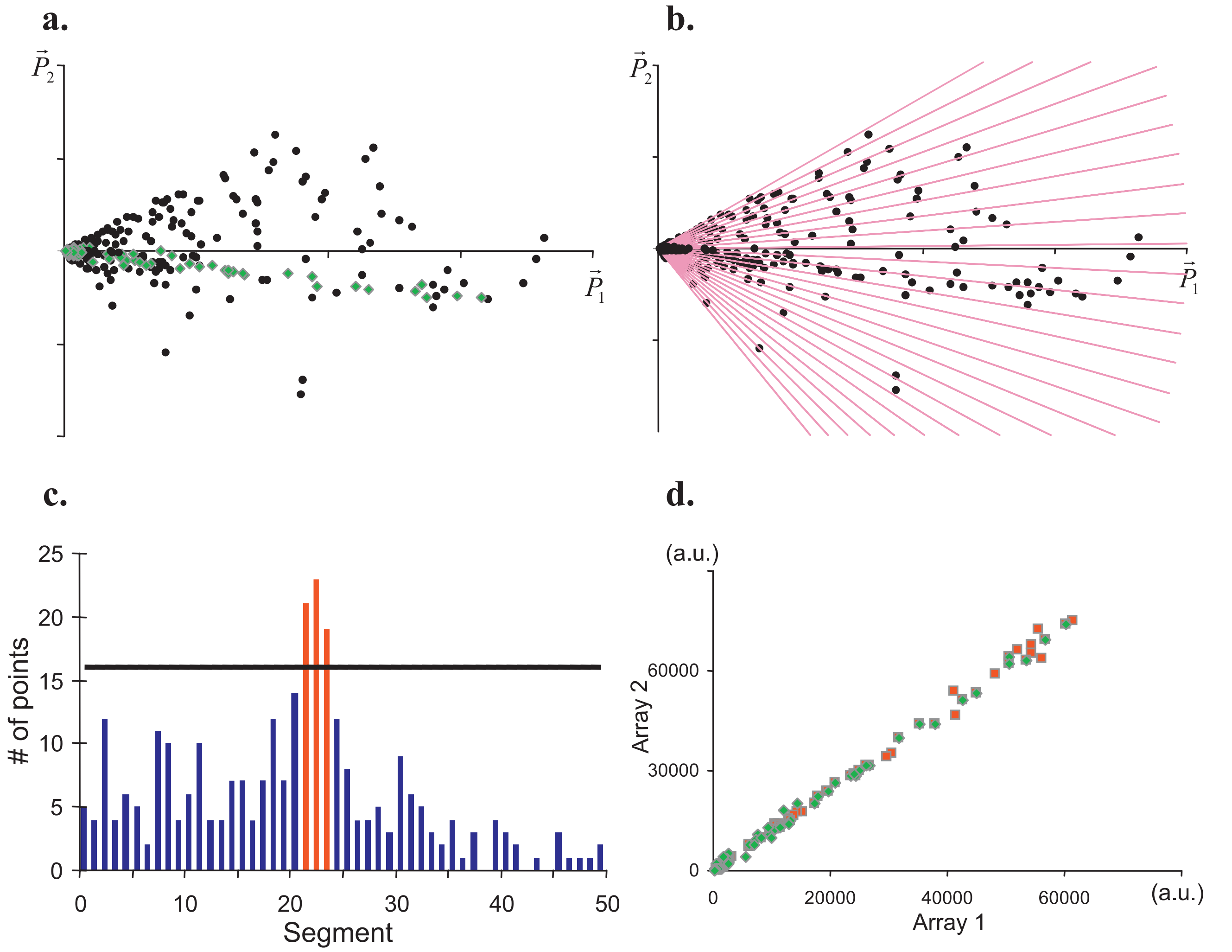

Fig. 4. (**a**) The data from Figure 3b, presented in the PC-plane; (**b**) schematic illustration of segmentation of the part of the PC-plane containing the data; (**c**) histogram of the angular densities of the segments; (**d**) 'true' (green) and PCA(line)-detected housekeeping genes (orange).

and # 65 [Fig. 6(b)]. The normalization lines [represented in orange and blue, respectively, for LR and PCA(line)] indicate that in the case of LR, a handful of strongly expressed genes are driving the normalization. A similar graph was obtained with arrays #1 and #58, which also showed a large difference between the two normalization procedures.

To determine how the number of segments in the plane impacts the estimated normalization coefficients, we ran the procedure with $p = 100$, 300, 400 and 500. In all cases, the procedure extracted essentially the same subset of normalizing housekeeping genes. The number of genes for each p was 1410, 1192, 1092 and 1162, respectively. We estimated a (5×5) correlation matrix of the derived normalization factors for each value of p. All coefficients in the correlation matrix were greater than 0.994, indicating the high degree of reproducibility between the derived normalization factors for different numbers of segments (p). We also estimated the coefficient of variation (COV) between the five series of estimates. The average COV for the 72 normalization factors was 1.71%.

DISCUSSION

Normalization of gene intensities in multi-array experiments is crucial for the ultimate biological interpretation to be meaningful (Hoffmann *et al*., 2002). Only after proper normalization can changes in expression of a given gene amongst the studied samples in the experiment be characterized quantitatively. Conversely, erroneous (or no) normalization may lead to inaccurate estimation of the changes in gene expression including wrong conclusions with regard to their up- or down-regulation. While optimal normalization is still a subject of discussion, individual investigators are faced daily with many questions about the analysis of these complex

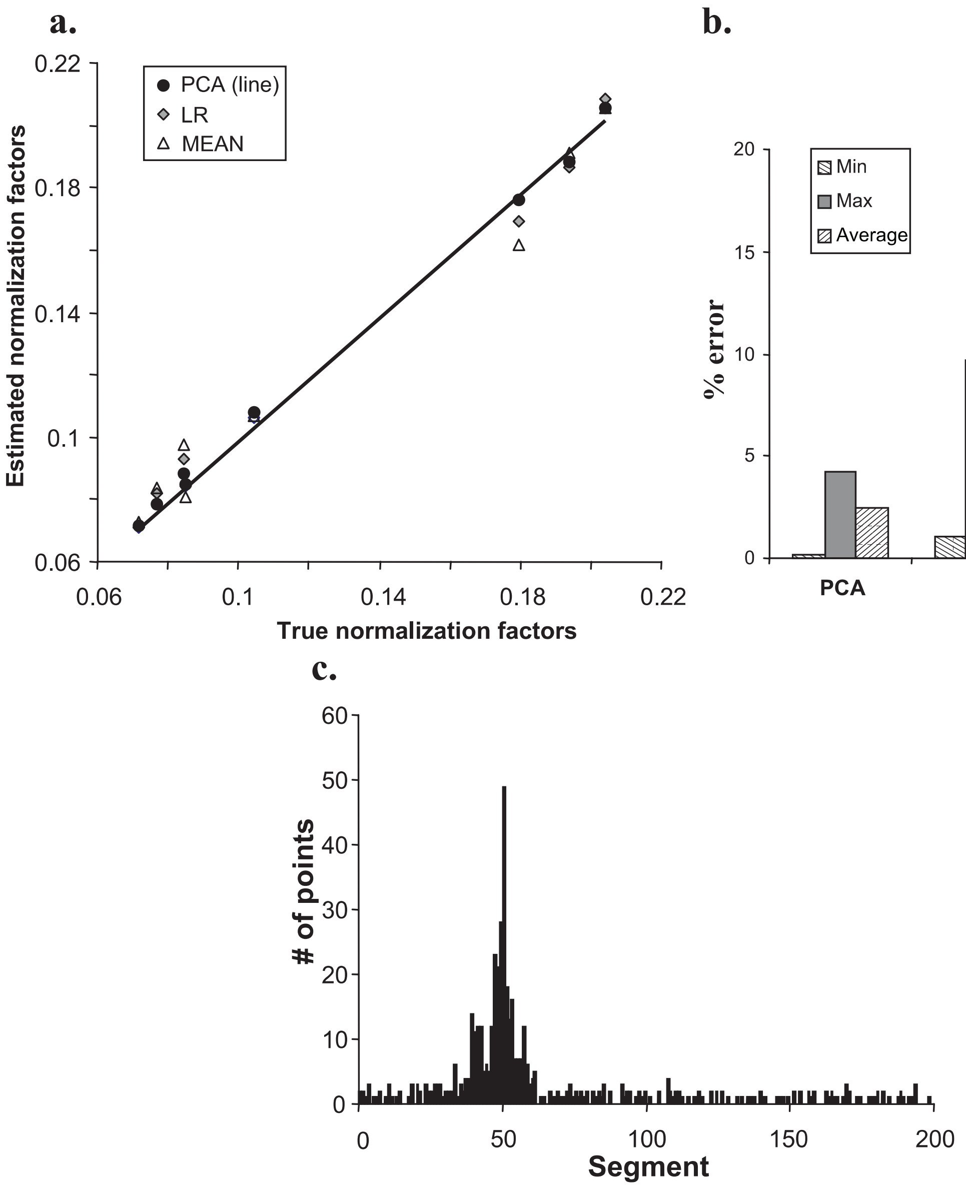

Fig. 5. (**a**) Relation of 'true' normalization factors and factors estimated via PCA(line), LR and MEAN in a simulated dataset containing eight arrays. The black line indicates the line of identity; (**b**) ranges (minimum and maximum) and average of the absolute values of relative errors of estimation of the normalization factors in the three estimates; (**c**) histogram of the angular densities of the segments in the PCA(line) for unfiltered data.

data. For example, should the array data be logarithmically transformed prior to normalization; should low intensity spots be discarded, and, if so, what is the right cut-off limit for this operation; should the mean or median intensity of the arrays be used for normalization; or alternatively, do 'designated' housekeeping genes play reliably their assigned role?

In this report, we address all these questions and present a simple procedure for normalization of datasets generated with single-channel arrays based on PCA. The procedure makes minimal assumptions about the data and does not require any pre-processing, pre-screening or filtering of the data.

The need for alternative normalization techniques arose with the realization that genes assumed as housekeeping and 'designated' by the manufacturers as such on arrays are not reliable for accurate data normalization. In the first experiment with HOSE cells, investigating a set of three cell lines with close genetic origin, the 'designated' housekeeping genes change in a coordinated fashion, and it is likely that they fulfill their role as normalizing genes. This result is anticipated

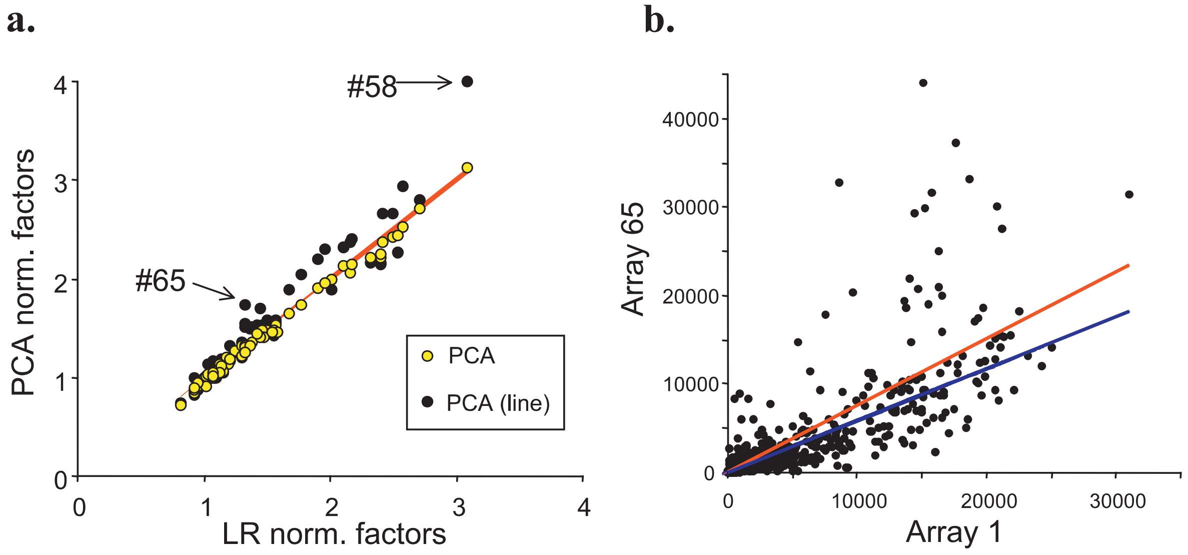

Fig. 6. (**a**) Correlation between LR- estimated (x-axis) and PCA- or PCA(line)-estimated (yellow series and black series, respectively) normalization factors for the LD. The orange line indicates the identity line. The arrows point at arrays with a large relative difference; (**b**) correlation plots of intensities of genes marked with 'P' in arrays #1 and #65. The normalization lines derived by the LR and PCA(line) estimates are indicated in orange and blue, respectively.

since the three cell lines were cultured under standard growth conditions and the observed differences in the global gene expression profiles are related to only a small subset of genes associated with the sequential transition of the cells through the process of malignant transformation. Conversely, in the second experiment, the 'designated' housekeeping genes appear to change differentially in response to treatment with Fenretinide. This is consistent with the dramatic biochemical changes associated with the process of cells undergoing programmed cell death (Querec, T.D. *et al.*, manuscript in preparation). The major alterations in the global gene expression profile that precedes and leads to the triggering of apoptosis affect the expression states of most housekeeping genes.

Pre-processing of the data prior to normalization is an important issue. Typical steps include background correction, logarithmic transformation and/or thresholding. We believe that the background should be removed prior to normalization, so that the normalization line goes through the origin. Although we simulated gene intensities, as described in the Materials and methods section, there is no theoretical basis to assume that real data comply with this distribution. Log-transformation has the advantage of transforming the noise distributions approximately to Gaussian. This property can be used for estimating the probabilities of differentially expressed genes (Kerr *et al.*, 2000). The PCA-based normalization procedure, however, is based on identifying the genes along the normalization line in the dataset and is invariant to prior transformation. Moreover, based on 'noise'-simulated data, as well as from the HOSE cell replicates, it is apparent that log-transformation may be detrimental to the analysis as it increases the relative contribution of the gene-independent noise in genes expressed at low levels. Because of these adverse effects, and the fact that by estimating the numbers of genes in the segmented plane the PCA(line) procedure allows low-expressed genes to be taken into consideration, we chose to implement our normalization procedure on raw (untransformed) data.

The described procedure is also insensitive with respect to prefiltering (thresholding) of the data, given that the parameter μ [Equation (2)] is adjusted appropriately. In the case of 'thresholded' data, $\mu = 1.96$ will be sufficient to discriminate between the sought housekeeping genes and the rest [Fig. 4(c)]. This μ-value will merely distinguish the 'noise' genes from the signal ones in non-prefiltered data. Thus, a larger μ [as in the case shown in Fig. 5(c)] is required to detect the normalizing genes sought. We therefore strongly recommend exploring the characteristics of the angular histogram of the data before setting the appropriate μ-value.

The only assumption made about the distribution of the intensities of the houseskeeping genes for PCA(line) is that they are distributed along a straight line. This assumption is very sensible for single-channel arrays, unlike the case of the double-channel arrays, where it is known that a nonlinear dependence exists between the gene expression levels among the two channels (Yang *et al.*, 2002). Furthermore, it has been shown recently that even for these arrays the linear and non-linear normalization methods perform similarly (Park *et al.*, 2003). In our experience, most of the nonlinear effects are due to improper scanning settings, which, besides the unwanted variations, produce saturated spots also.

We consider the identification of the housekeeping genes with intensities within the linear range, as proposed by the PCA(line) routine, to be a reliable and robust source for normalization.

The linearity is the basis of the stability of the approach with respect to the parameter p—it is sufficient to detect a small subset of S to identify uniquely the normalization line. Conversely, a larger set of genes along this line will not impede the calculation of the normalization parameters. Still, in order to obtain meaningful histograms of the number of genes in each segment, we recommend that p initially be selected to contain on average at least 10 genes per segment. The condition for linearity naturally excludes genes with saturated expression levels and it thus contributes significantly to reducing the interference of these typically large signals in the normalization process.

Conditions (1) and (2) (see Introduction) are instrumental for the successful performance of the referenced normalization procedures. However, in single-channel arrays, such as the Affymetrix platform and radiolabeled filter arrays, it is a common phenomenon that the detected number of up-regulated genes is larger than the number of the down-regulated ones. This is due to the fact that the signals of genes expressed at low levels and undergoing down-regulation are close to or below the background level, and, therefore, their change is either undetected or deemed statistically insignificant. When these conditions hold, as in the case of the simulated data in Figure 3(a), PCA will be successful in determining the normalization factors with the following advantages, as compared with the other referenced techniques:

- It provides an objective measure through the magnitude of the first eigenvalue of how ‘tightly’ the data are distributed along the first PC.
- It simultaneously determines normalizing coefficients for the entire dataset. A common approach for normalization of multiple experiments is to choose one array as the baseline and to apply normalization (Golub *et al*., 1999). In order to avoid the lack of symmetry of this procedure, the baseline is computed frequently as the average gene expression profile (Tusher *et al*., 2001). This is achieved naturally with PCA as the first PC is an approximation of the ‘average’ array in the dataset.
- Viewing the entire set of multiple array data simultaneously allows proper down-weighing of the ‘noise’ genes, which, during individual comparisons, may affect strongly the calculation of the normalization coefficients.

The advantages of PCA are underscored in the LD example, where a single PCA step applied to the entire dataset estimates normalization coefficients that are almost identical to the ones determined by the pair-wise LR procedures, using only well measured genes in each pair [Fig. 6(a)].

The PCA(line) procedure, besides having the above listed general advantages of PCA, can also deal successfully with situations where conditions (1) and (2) do not apply. In the simulated datasets, the PCA(line) results are closest to the true values as judged by the relative mean-square errors from the three procedures tried. Visual inspection of the LR and PCA(line) normalization lines in the graph shown in Figure 6(b) suggests that this is also true for the Affymetrix data. In addition, it eliminates the need for using a baseline array, which, as shown by Bolstad *et al*. (2003), has a clear disadvantage relative to the complete data methods for normalization such as the one proposed here.

In conclusion, the proposed normalization procedure improves significantly the accuracy and precision of the measured gene expression levels. Such procedures will become even more relevant with further refinement and standardization of the microarray technology.

ACKNOWLEDGEMENTS

The authors would like to thank Dr S. Litwin for reviewing the manuscript critically and for his suggestions for improving it further. The authors would also like to thank Dr P. Tamayo for his helpful discussion regarding the Affymetrix array dataset. The work described in this report was supported by funds provided through NIH grant R29-CA73676 to C.P., P01-CA41078 to T.R.B., P50-CA83638 (PI: R. Ozols) and a Guzik Foundation Award to C.P. and R.S. C.P. is a Liz Tilberis Scholar of the OCRF, Inc.

REFERENCES

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci., USA*, **97**, 10101–10106.

Alter,O., Brown,P.O. and Botstein,D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl Acad. Sci., USA*, **100**, 3351–3356.

Anderson,T.W. (1971) *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bowtell,D.D. (1999) Options available—from start to finish—for obtaining expression data by microarray. *Nat. Genet.*, **21**, 25–32.

Butte,A.J., Dzau,V.J. and Glueck,S.B. (2001) Further defining housekeeping, or ‘maintenance,’ genes Focus on ‘A compendium of gene expression in normal human tissues’. *Physiol. Genomics*, **7**, 95–96.

Debouck,C. and Goodfellow,P.N. (1999) DNA microarrays in drug discovery and development. *Nat. Genet.*, **21**, 48–50.

Duggan,D.J., Bittner,M., Chen,Y., Meltzer,P. and Trent,J.M. (1999) Expression profiling using cDNA microarrays. *Nat. Genet.*, **21**, 10–14.

Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**, 179–205.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hartemink,A., Gifford,D., Jaakola,T. and Young,R. (2001) Maximum likelihood estimation of optimal scaling factors for expression array normalization. *Proc. SPIE*, **4266**, 132–140.

Hoffmann,R., Seidl,T. and Dugas,M. (2002) Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome Biol.*, **3**, RESEARCH0033.

Kepler,T.B., Crosby,L. and Morgan,K.T. (2002) Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol.*, **3**, RESEARCH0037.

Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Lander,E.S. (1999) Array of hope. *Nat. Genet.*, **21**, 3–4.

Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci., USA*, **98**, 31–36.

Misra,J., Schmitt,W., Hwang,D., Hsiao,L.L., Gullans,S. and Stephanopoulos,G. (2002) Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res.*, **12**, 1112–1120.

Moon,R.C., Thompson,H.J., Becci,P.J., Grubbs,C.J., Gander,R.J., Newton,D.L., Smith,J.M., Phillips,S.L., Henderson,W.R., Mullen,L.T., Brown,C.C. and Sporn,M.B. (1979) N-(4-hydroxyphenyl)retinamide, a new retinoid for prevention of breast cancer in the rat. *Cancer Res.*, **39**, 1339–1346.

Nielsen,T.O., West,R.B., Linn,S.C., Alter,O., Knowling,M.A., O'Connell,J.X., Zhu,S., Fero,M., Sherlock,G., Pollack,J.R. *et al*. (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, **359**, 1301–1307.

Park,T., Yi,S.G., Kang,S.H., Lee,S., Lee,Y.S. and Simon,R. (2003) Evaluation of normalization methods for microarray Data. *BMC Bioinformatics*, **4**, 33.

Patriotis,P.C., Querec,T.D., Gruver,B.N., Brown,T.R. and Patriotis,C. (2001) ArrayExplorer, a program in visual basic for robust and accurate filter cDNA array analysis. *Biotechniques*, **31**, 862–872.

Peterson,L.E. (2003) Partitioning large-sample microarray-based gene expression profiles using principal components analysis. *Comput. Methods Programs Biomed.*, **70**, 107–119.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.*, **2**, 418–427.

Sapir,M. and Churchill,G.A. (2000). Published: The Jackson Laboratory **Poster**.

Schadt,E.E., Li,C., Ellis,B. and Wong,W.H. (2001) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell Biochem. Suppl.*, **37**(suppl.), 120–125.

Schadt,E.E., Li,C., Su,C. and Wong,W.H. (2000) Analyzing high-density oligonucleotide gene expression array data. *J. Cell Biochem.*, **80**, 192–202.

Schuchhardt,J., Beule,D., Malik,A., Wolski,E., Eickhoff,H., Lehrach,H. and Herzel,H. (2000) Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, **28**, E47.

Selvey,S., Thompson,E.W., Matthaei,K., Lea,R.A., Irving,M.G. and Griffiths,L.R. (2001) Beta-actin—an unsuitable internal control for RT–PCR. *Mol. Cell Probes*, **15**, 307–311.

Stoyanova,R. and Brown,T.R. (2002) NMR spectral quantitation by principal component analysis. III. A generalized procedure for determination of lineshape variations. *J. Magn. Reson.*, **154**, 163–175.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci., USA*, **98**, 5116–5121.

Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Zien,A., Aigner,T., Zimmer,R. and Lengauer,T. (2001) Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, **17** (Suppl. 1), S323–S331.

APPENDIX 1: ALGORITHM DESCRIPTION

(1) Construct the data matrix $\mathbf{D}(i, j)$, where

$$
\begin{aligned}
&i = 1, \ldots, n(n\text{—total number of genes on each array}),\\
&j = 1, \ldots, m(m\text{—total number of arrays in the}\\
&\qquad\qquad \text{dataset}).
\end{aligned}
$$

(2) (Optional) thresholding of the data:

(2.1) Set the values in $\mathbf{D}$ smaller than a given value (e.g. 200 a.u. for the Clontech data) to 0.

(2.2) Remove from $\mathbf{D}$ genes with 0 intensities in at least one array, resulting in a new data matrix $\mathbf{D}'(n' \times m)$, where $n' \leq n$.

(3) PCA of $\mathbf{D}$ (here and in the rest of the text $\mathbf{D}$ should be substituted by $\mathbf{D}'$ in the case of thresholding, as well as n by n').

(3.1) Calculate $\mathbf{C}$—the covariance matrix of $\mathbf{D}$:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{D}^{\mathbf{T}}\mathbf{D},$$

where $\mathbf{D}^{\mathbf{T}}$ denotes the transpose matrix of $\mathbf{D}$.

(3.2) Calculate eigenvectors $\mathbf{Q}$ and eigenvalues $\boldsymbol{\Lambda}$ of the covariance matrix $\mathbf{C}$, i.e.:

$$\mathbf{CQ} = \mathbf{Q}\boldsymbol{\Lambda}$$

The rows in $\mathbf{Q}$ are the PCs $\vec{P}_1, \vec{P}_2, \ldots, \vec{P}_m$.

(3.3) Calculate the scores $R = \mathbf{D}\,\mathrm{P}^{\mathrm{T}}$.

(4) Let R_1^i and R_2^i be the scores of the i-th gene along $\vec{P}_1$ and $\vec{P}_2$.

(4.1) Disregard genes for which $R_2^i = 0$.

(4.2) Calculate the angle $\varphi_i, i = 1, \ldots, n$ (in radians), between $\vec{P}_2$ and the vector with coordinates (R_1^i, R_2^i), as follows:

$$\varphi_i = \begin{cases} 2\pi + \arctan(R_1^i/R_2^i), \\ \quad \text{if } R_1^i \le 0 \text{ and } R_2^i > 0, \\ \arctan(R_1^i/R_2^i) \\ \quad \text{if } R_1^i > 0 \text{ and } R_2^i > 0, \\ \pi + \arctan(R_1^i/R_2^i) \\ \quad \text{if } R_1^i > 0 \text{ and } R_2^i < 0, \end{cases} \qquad i = 1, \ldots, n.$$

(5) Segment the part of the plane defined by the first 2 PCs in p partitions.

(5.1) Determine the segment $\theta = \max(\varphi_i) - \min(\varphi_i)$

(5.2) Determine a step $\delta = \theta/p$

(5.3) Define the subset of genes s_k in each of the p segments, defined as

$$s_k \in [(k-1)\delta \min(\varphi_i), k\delta \min(\varphi_i)], \\ k = 1, \ldots, p.$$

(6) Determine the subset of housekeeping genes **S**.

(6.1) Determine the number of genes θ_k in each subset s_k.

(6.2) Estimate the mean $M(\theta_k)$, and variance, $V(\theta_k)$, of the distribution of θ_k.

(6.3) Evaluate if

$$\theta_k > M(\theta_k) + \mu\sqrt{V(\theta_k)}$$

holds for any k. μ is a cut-off parameter, which can be set to 1.96 if a normal distribution of θ_k is assumed [see body of the paper, Equation (2)].

If none of the segments satisfies the condition it means that either none of the genes can serve as a housekeeping gene ($\mathbf{S} \equiv \emptyset$) or all genes in the dataset can be assumed to be housekeeping genes ($\mathbf{S} \equiv \mathbf{D}$). Then the loadings of $\vec{P}_1$ (3.2) may be used as normalizing factors.

(6.4) The expression levels of the genes in each array should be divided by these loadings.

End of the Procedure

(6.5) Let Z denote the set of these segments that satisfy the condition in 6.3. If for a certain q, $\zeta_q \in Z$, then

(6.5.1) If $\zeta_{q+1} \notin Z$, then

(6.5.1.1) If there are no other qs, for which $\zeta_q \in Z$, then proceed as in 6.4.

(6.5.1.2) Conversely, proceed as in 6.5.

(6.5.2) If $\zeta_{q+1} \in Z$, then the genes in these two segments are assumed to be housekeeping genes; $\mathbf{S} \equiv s_q \cup s_{q+1}$. Add to S the genes of any consecutive segments that belong to Z.

(6.5.2.1) Apply PCA (3.2) to the gene expression levels in **S**. The loadings of $\vec{P}_1$ can be used as normalizing factors. The expression levels of the genes in each array should be divided by these loadings.

End of the Procedure

APPENDIX 2: SIMULATED DATASET

Let g_{i1} be the gene intensity of the i-th gene in the first array ($i = 1, 2, \ldots, 500$). The corresponding intensities in the second array in SD1 were generated as follows.

$$\left| \begin{array}{ll} g_{i2} = q_{12} * \min[\alpha_{\text{up}}\, g_{i1}, \beta_{\text{up}}] & i = 1, \ldots, 200, \\ g_{i2} = q_{12} * \max[\alpha_{\text{down}}\, g_{i1}, \beta_{\text{down}}] & i = 201, \ldots, 350, \\ g_{i2} = q_{12} * g_{i1} & i = 351, \ldots, 500, \end{array} \right. \tag{A.1}$$

where $q_{12} = 1.2$, and the αs and βs are random numbers within the following intervals:

$$\begin{aligned} &\alpha_{\text{up}} = (1, 10], \\ &\beta_{\text{up}} = (g_{i2}, g_{\max}], \quad \text{where } g_{\max} = 80\,000, \\ &\alpha_{\text{down}} = (0, 1/10], \\ &\beta_{\text{down}} = (g_{\min}, g_{i2}], \quad \text{where } g_{\min} = 0. \end{aligned}$$

APPENDIX 3: SIMULATED DATASET

Let g_{ij} be the gene intensity of the i-th gene in the j-th array ($i = 1, 2, \ldots, 500$; $j = 1, 2, \ldots, 7$). Equation (A.1) describes the generation of the data in SD2 (q_{12} substituted correspondingly with q_{1j}, randomly generated scaling parameters between 0.3 and 3), derived from the intensities of the genes in the first array, where α_{up}^j and α_{down}^j are consistent with a simulated gradual increase in fold of changes between 1.5 and 4.5 with an increment of 0.5, both for up- and down-regulated genes. Formally,

$$\begin{aligned} &\alpha_{\text{up}}^j = (1, 1 + j * \text{step}], \\ &\alpha_{\text{down}}^j = (0, 1/(1 + j * \text{step})], \end{aligned} \qquad j = 1, \ldots, 7$$

where step $= 0.5$.